# 12 COMPUTERIZED ADAPTIVE TESTING: THEORY, APPLICATIONS, AND STANDARDS

Ronald K. Hambleton
Jac N. Zaal
Jo P. M. Pieters

The testing industry in the United States has undergone major changes in the last 15 years. Classical test theory models and methods which have been in wide use for 60 years or more are being replaced by new test theories and methods, most notably item response theory (Hambleton & Swaminathan, 1985; and Weiss & Yoes, in Chapter 3). For example, many of the important standardized achievement and aptitude tests (e.g., Scholastic Aptitude Test, California Achievement Tests, the Stanford Achievement Tests, and the Woodcock-Johnson Psycho-Educational Battery) are developed using item response model principles and procedures.

A second major testing change in the United States has been the wide interest in and great use of criterion-referenced tests, sometimes called objectives-referenced tests, basic skills tests, competency tests, or credentialing exams (see, for example, Berk, 1984; and Hambleton & Rogers, Chapter 1). These tests provide for score interpretations in relation to well-defined bodies of content and standards of performance. These tests are being used in the United States for (1) awarding high school diplomas, (2) monitoring student progress within a grade and from one grade to the next, (3) awarding certificates or licenses to professionals in over 900 professions, and (4) determining job competence in the armed services and industry. There are many other uses for these tests as well.

Numerous changes in testing are also taking place at the present time



R. K. Hambleton et al. (eds.), *Advances in Educational and Psychological Testing: Theory and Applications*

because of the availability of new computer capabilities (see, for example, Brzezinski & Hiscox, 1984; Bunderson, Inouye, & Olsen, 1989; Wainer, 1988). There is now wide recognition of the computer's immense power for storing test information (e.g., test items), and for producing, administering, and/or scoring tests. Many testing agencies in the United States already store their item pools in computer memory. These test items can be recalled on an as-needed basis. Computer software is also available now for tests to be completely constructed to fit a set of psychometric specifications (see, for example, Assessment Systems Corporation, 1988; WICAT Systems, 1986). Especially attractive is the ability to prepare tests for printing without any additional typing or typesetting of test items. Errors in test production can be eliminated because of this capability.

Presently, computers are also being used in some testing programs (see, for example, experimental versions of the *Differential Aptitude Tests*) to "tailor" the particular items an examinee is administered in a test. In this way, tests can be shortened without any loss of measurement precision. After an examinee responds to a set of test items (sometimes only one item) presented at a computer terminal, an initial ability estimate for the examinee is obtained. The computer is programmed to select the next set of items from the available item bank for administration which will contribute the most information to the estimation of the examinee's ability. Some details for how test items are selected, and ability estimates are obtained, will be addressed later in this chapter. The administration of items to the examinee continues until some specified number of items is administered or a desired level of measurement precision is obtained.

Weiss (1985) and Reckase (1989) and others have drawn attention to the fact that the earliest application of tailored or adaptive testing goes back to the work of Binet on intelligence testing in 1908, so the idea is not a new one in testing. Still, little additional work took place on adaptive testing until Fred Lord at the Educational Testing Service began a long comprehensive research program beginning in the late 1960s (for a review of his work, see Lord, 1980). Lord's motivation for pursuing adaptive testing was that he felt fixed length tests were inefficient for many examinees, especially low- and high-ability examinees. With the power of the computer to select, present, and score test items, he felt that tests could be shortened without any loss of measurement precision if the test items administered to each examinee were chosen to provide maximum information about the examinee's ability. In theory, each examinee would be administered a unique set of items. Since the late 1960s, a substantial amount of research has been supported by the U.S. Armed Services, the U.S. Civil Service Commission, and other federal agencies, special con-

ferences have been held, and the number of published papers on adaptive testing is in the hundreds (see, for example, Wainer, 1990; Weiss, 1983).

However, despite the promise of computers for facilitating test development, and opening up new possibilities for test administrations, item formats, and scoring, considerably more research is needed before we will know the full benefits, as well as shortcomings, of computer-administered achievement and aptitude tests. For comprehensive reviews of the roles and problems of computer use with testing, readers are referred to Brzezinski and Hiscox (1984) and Bunderson, Inouye, and Olsen (1989). This chapter will be focused more narrowly on the uses of computers in implementing adaptive testing strategies. Specifically, the purposes of this chapter are (1) to introduce the measurement models that underlie the applications of adaptive testing, (2) to highlight several promising adaptive testing applications, and (3) to review some developed guidelines by Green and his associates (1982, 1984) for designing and using adaptive tests.

## Introduction to Item Response Theory

With few exceptions, testing is carried out in settings in which a group of individuals take the same test (or parallel forms). Typically, individuals in the group will vary in the ability being measured by the test. It can be shown easily that the test would measure maximally the ability of each individual in the group if test items were presented to each individual such that the probability of answering each item correctly was around .50, or around .60 if guessing is possible (see, for example, Hambleton & Swaminathan, 1985). Since, in general, examinees differ in their ability levels, higher-ability examinees would need to be administered relatively harder items and lower-ability examinees would need to be administered relatively easier items in order that both groups might have (approximately) 50 percent probabilities of answering correctly the test items administered. This, of course, is not possible using a single test; consequently, there is a need for "tailored tests" or "adaptive testing" (Green, 1983; Lord, 1971, 1980; Weiss, 1982, 1983) if testing time is to be reduced without loss of measurement precision.

Item response models are particularly important in adaptive testing because it is possible to derive, using item response models, ability estimates that are independent of the particular choice of test items administered. Thus, examinee abilities can be compared even though the examinees may have taken sets of test items of varying difficulty (Urry, 1977, 1983). And,

in theory, each examinee will receive a different set of test items. In fact, a primary reason for Fred Lord switching his research focus from true score theory to item response theory in the late 1960s was to develop both the theoretical and practical bases for estimating ability independent of the selection of test items. Such a feature was not possible within a true score model framework.

In a few words, item response theory (IRT) postulates that (1) examinee test performance can be predicted (or explained) by a set of factors called traits, latent traits, or abilities, and (2) the relationship between examinee item performance and the set of traits assumed to be influencing item performance can be described by a monotonically increasing function called an item characteristic function. This function specifies that examinees with higher scores on the traits have higher expected probabilities for answering an item correctly than examinees with lower scores on the traits.

In applying item response theory to measurement problems, a common assumption is that there is one dominant factor or ability which can account for item performance. This assumption is made, for example, in nearly all of the current applications of adaptive testing. In the one-trait or one-dimensional model, the item characteristic function is called an item characteristic curve (ICC) and it provides the probability of examinees answering an item correctly for examinees at different points on the ability scale defined for the trait measured by the test (Hambleton & Swaminathan, 1985; Lord, 1980; and Weiss & Yoes, Chapter 3).

Corresponding to each item also is an item information function, the shape of which depends on the item statistics. The important feature of an item information function is that it is defined over the same scale on which ability is measured and indicates the contribution of the item to ability estimation at points along the ability continuum. It is this feature which is utilized in item selection. Items are preferred which provide the most information about examinee ability which can be determined once an ability estimate for the examinee is available. Items providing the most information are, generally, the same ones where the examinee has an (approximately) 50 percent chance of answering correctly though the situation is more complicated when items vary substantially in their item parameter values.

In addition to the assumption of test unidimensionality, it is common to assume that the item characteristic curves are described by one, two, or three parameters. The specification of the mathematical form of the ICCs and the corresponding number of parameters needed to describe the curves determines the particular item response model.

In any successful application of item response theory, parameter

estimates corresponding to the choice of ICCs are obtained to describe the test items, and ability estimates are obtained to describe the performance of examinees. Also, successful applications require that there be evidence that the chosen item response model, at least to an adequate degree, fits the test data set (Hambleton & Swaminathan, 1985). The three-parameter logistic model is probably the model of choice by current CAT-advocates (Green & associates, 1984; Lord, 1980; Weiss, 1983). The most important reason is that a well-fitting test model is essential to the success of any CAT-applications and the three-parameter model fits test data better than either the one- or two-parameter models.

Item response theory has become a very popular topic for research in the measurement field. Numerous IRT research studies have been published in the measurement journals, many conference presentations have been made, and applications of the theory have been made to many pressing measurement problems (i.e., test score equating, study of item bias, test development, item banking, and adaptive testing) in the last several years (see, for example, Hambleton, 1983, 1989; Hambleton & Swaminathan, 1985; Lord, 1980; Weiss, 1983). Interest in item response theory stems from two desirable features which are obtained when an item response model fits a test data set: descriptors of test items (item statistics) are not dependent upon the particular sample of examinees from the population of examinees for whom the test items are intended, and the expected examinee ability scores do not depend upon the particular choice of items from the total pool of test items to which the item response model has been applied. Invariant item and examinee ability parameters, as they are called, are of immense value to measurement specialists. Neither desirable feature is obtained through the use of classical test models. The first feature removes some of the difficulties associated with obtaining statistics for a large bank of test items that is needed for adaptive testing. Sample characteristics are less important and consequently new items can be field-tested and added to the item bank on an as-needed basis. The second feature permits meaningful comparisons of examinees on the ability measured by the adaptive test, though examinees will have been administered statistically unequal tests.

Today, item response theory is being used by most of the large U.S. test publishers, state departments of education, and industrial and professional organizations, to construct norm-referenced and criterion-referenced tests, to investigate item bias, to equate tests, and to report test score information. Some issues and technical problems remain to be solved in the IRT field but it would seem that item response model technology is more than adequate at this time to serve a variety of uses, notably adaptive testing

(see, for example, Lord, 1980; Wainer, 1990; Weiss, 1982, 1983, 1985). IRT is especially relevant for adaptive testing because ability estimates can be compared even though examinees take different items. Also, optimal item selection can be done because the usefulness of items for a given ability level can be determined.

## Adaptive Testing

In adaptive testing, an attempt is made to match the difficulties of the test items to the ability of the examinee being measured. To match test items to ability levels requires a large pool of items whose statistical characteristics are known so that suitable items may be drawn (Millman & Arter, 1984). Since the item selection procedure does not lend itself easily to paper-and-pencil tests, the adaptive testing process is typically done by computer. (Exceptions to this rule are presented in the work of Lord, 1971.) According to Lord (1980), a computer must be programmed to accomplish the following in order to tailor a test to an examinee:

1. Predict from the examinee's previous responses how the examinee would respond to various test items not yet administered.
2. Make effective use of this knowledge to select the test item to be administered next.
3. Assign at the end of testing a numerical score that represents the ability of the examinee tested.

Advantages of computer adaptive testing (CAT), in addition to shortening tests without loss of measurement precision, include enhanced test security, testing on demand, answer sheets are not needed, test pace is keyed to the individual, quick test scoring and reporting, minimizing test frustration for some examinees, greater standardization, removal of "defective items" from the item bank when they are identified, greater flexibility in item formats and minimizing test administration supervision time (Olsen & associates, 1989).

The research on adaptive testing to date has been focused on six components:

1. The choice of IRT model,
2. The item bank,
3. Choice of starting point,
4. Selection of test items,

5. Scoring/ability estimation, and
6. Choice of stopping rule.

Each of these components will be briefly considered next.

### *The Choice of IRT Model*

This component is important because the validity of the ability estimates depends upon the fit of the chosen model to the test data. Model selection was considered at some length in the Weiss and Yoes chapter. Most of the adaptive testing projects to date have used the three-parameter logistic model.

### *The Item Bank*

Adaptive testing, like any fixed-length testing, requires a valid set of test items. The one main difference is that effective adaptive testing requires the presence of a large as well as valid item bank. Of special interest are highly discriminating items since these items are especially useful in producing short tests with optimal measurement characteristics. Large banks are needed to minimize problems associated with test security and to provide the basis for producing optimal tests for examinees at a wide range of abilities. A second difference is that "defective items" are apt to cause more problems in adaptive testing because these tests are relatively short. Therefore, considerable effort must be expended to insure high quality test items.

Of course, IRT item statistics are needed, too, which are obtained through properly designed field tests involving large samples of examinees. As part of the field-testing, studies of unidimensionality are normally carried out, and items which do not meet this assumption to a reasonable degree must be deleted. In time, multi-dimensional IRT models may be operational in adaptive testing (Reckase, 1989), but, at the moment, these models are new and the amount of research on them has been low.

### *Choice of Starting Point*

One of the factors that influences the number of items in an adaptive test is the starting place. A good starting place would probably involve admini-

stering items that are matched to the examinee's ability level. In the absence of any information about the examinee, normally an item of average difficulty is chosen. However, information about the examinee's ability level such as might be inferred from background educational data or self-reports can be helpful. Most researchers seem pretty sanguine about starting points, feeling that, as long as the test is not too short, the starting place for testing doesn't make much difference. About the only opposition on this point comes from Wainer and Kiely (1987) who feel that test anxiety and frustration are increased with inappropriate starting points.

### *Selection of Test Items*

Research has been done on a variety of adaptive testing strategies built on the following decision rule: if an examinee answers an item correctly, the next item should be more difficult; if an examinee answers incorrectly, the next item should be easier. These strategies can be broken down into *two-stage strategies* and *multistage strategies*. The multistage strategies are either of the *fixed branching* variety or the *variable branching* variety.

In the two-stage procedure (Lord, 1980), all examinees take a routing test and, based upon this test, are directed to one of a number of tests constructed to provide maximum information at certain points along the ability continuum. Ability estimates are then derived from a combination of scores from the routing test and the optimum test. The two-stage strategy has been popular with some test developers because it is possible to implement without computers and in group test administrations. Normally, the short routing test is scored by hand, and then, based upon the examinee's score, a second and substantially longer test is assigned which is roughly matched to the examinee ability level. In practice, usually three to five tests are available for assignment following the routing test, though certainly more tests could be used. Perhaps because of the modest reliability of the routing test (due to its shortness), assignment of students to more than five tests does not seem warranted.

Whereas the two-stage strategy requires only one branching step from the routing to the optimum test, multistage strategies involve a branching decision after the examinee response to each item. If the same item structure is used for all individuals, but each individual can move through the structure in a unique way, then it is called a fixed-branching model. The question of how much item difficulty should vary from item to item leads to considerations of structures with constant, decreasing, or increasing step size. One criticism of fixed branching schemes is that they usually

only consider one item statistic, item difficulty, though there are other characteristics of the item, e.g., discriminating power, that influence examinee performance, too.

For these multistage fixed-branching models, all examinees start at an item of median difficulty and, based upon a correct or an incorrect response, pass through a set of items that have been arranged in order of item difficulty. After having completed a fixed set of items, either of two scores is used to obtain an estimate of ability: the difficulty of the (hypothetical) item that would have been administered after the *n*th (last) item, or the average of the item difficulties, excluding the first item and including the hypothetical $n$ + first item.

Other examples of fixed multistage strategies include the flexi-level test and the stratified-adaptive (stradaptive) test (Lord, 1971; Weiss, 1982). The flexi-level test, which can be represented in a modified pyramidal form, has only one item at each difficulty level. The decision rule for using this test is: following a correct response, the next item given is the item next higher in difficulty that has not been administered. Following an incorrect response, the item next lower in difficulty that has not been administered is given. The stradaptive test, on the other hand, has items stratified into levels according to their difficulties. Branching then occurs by difficulty level across strata and can follow any of a number of possible branching schemes.

The variable-branching structures are multistage strategies that do not operate with a fixed item structure. Rather, at each stage of the process, an item in the established item bank is selected for a certain examinee in a fashion such that the item, if administered, will maximally reduce the uncertainty of the examinee's ability estimate. After administration of the item, the ability estimate is either re-calculated using maximum likelihood procedures (Lord, 1980) or Bayesian procedures.

There are two main methods currently for item selection in variable-branching structures when conducted within an IRT framework (Kingsbury & Zara, 1989). The first, *maximum information* (Weiss, 1982), involves the selection of items which provide maximum information at the current estimate of the examinee's ability in the testing process. To avoid the same items being selected time and time again (items with the highest levels of discriminating power, in general provide the most information) and thereby (possibly) affecting test security and, subsequently, test validity, Green and associates (1982) have suggested that items be selected on a random basis from *among* those items that provide the most information at the ability level of interest.

The second method, *Bayesian item selection* (Owen, 1975), involves the

selection of test items which contribute most to minimizing the "posterior belief distribution" about an examinee's ability based upon the knowledge of whether or not the examinee answered the item correctly. The method sounds complicated but it is quite straightforward to apply in practice. Like all Bayesian methods, however, the success of the method depends on the appropriateness of the initial or prior estimate of examinee ability. When the priors are not realistic, or especially when the tests are short, bias in the ability estimates can occur.

### *Scoring/Ability Estimation*

A distinct advantage of computerized adaptive testing is that test scoring/ability estimation is being carried out while the test is being administered, thus facilitating quick feedback of results to examinees. There are two primary IRT methods of ability estimation, *maximum likelihood* (Lord, 1980) and *Bayesian* (Weiss, 1982). Both methods work fine, though maximum likelihood can be a problem with small numbers of test items (estimates are not possible with scores of 0 or n, where n is the number of items administered, for example), and Bayesian estimates tend to be biased.

### *Choice of Stopping Rule*

One of the keys to the success of adaptive testing is knowing when to terminate the testing. Several methods and combinations of methods are currently used. In one, testing is continued until some acceptable level of measurement error is achieved. In this way, ability estimates are all at the same level of measurement when testing is terminated (this parallels measurement within a classical test theory framework) though the number of items administered to each examinee will vary. It would also be possible to specify some acceptable but unequal levels of measurement precision for different ability levels. For example, a decision could be made that more precision is needed with middle abilities than for those at the extremes.

Another method involves setting a fixed number (not too large) of test items for the set of examinees. Testing time is (approximately) constant for all examinees, but the standard error of ability estimation will vary from one examinee to the next. In some applications, a minimum number of items which must be administered is specified, and then testing is continued until the measurement error associated with the ability estimate attains some prespecified acceptable level. This method often adds credibility to

the testing in the minds of the examinees. Short tests are often viewed suspiciously by examinees. By specifying a minimum number of test items, some of the criticism which might result from administering a very short test is avoided.

In practice, too, sometimes an upper-bound on the number of items that can be administered is set. In this way, examinees and computers need not be tied up for unrealistic periods of time.

One interesting variation on the above methods arises in the context of criterion-referenced testing (Weiss & Kingsbury, 1984). Here, a cut-off score is available to separate examinees into "mastery" and "nonmastery" states. Testing can continue until the probability of assigning the examinee to the correct mastery state exceeds some acceptable level (e.g., 90%). Or, alternately, testing can continue until the confidence band around the examinee's ability estimate is on one side or the other of the cut-off score.

### *Evaluation*

There are a number of ways in which items can be tailored to ability, as well as ways of computing ability estimates. What is needed, however, is a mechanism for evaluating the results of studies obtained from these various procedures. The mechanism for evaluation should not be based on group statistics such as correlation coefficients because the crux of the problem is to determine the accuracy with which ability can be estimated for a single examinee. Almost all the comparative studies in the literature have compared tests constructed using various procedures by making use of test information functions (see, for example, Weiss, 1982). Does adaptive testing work? The evidence, simulated and live-testing, is substantial that it does. Readers are referred to Moreno and associates (1984), Ward, Kline, and Flaugher (1986), Weiss (1982, 1985), Wainer (1990), and Weiss and Kingsbury (1984) for some of this evidence. Adaptive testing procedures provide more information at the extremes of the ability distribution than do any of the standard tests used for comparative purposes and they provide adequate information at medium-difficulty and medium-ability levels (where standard tests cannot be surpassed).

### *Summary*

In the United States in the last three or four years there have been numerous applications of computer-adaptive tests. The U.S. Armed Services, for

example, is planning to administer the Armed Services Vocational Aptitude Battery using computer-adaptive testing procedures, and one testing company currently has adaptive testing projects in over 180 school districts (see Olsen & associates, 1989). In addition, most of the major testing firms in the United States are researching possible uses of computer-adaptive testing. The next few years should see many applications along with evaluative data concerning the success of these applications.

## Promising Applications

In this section, two adaptive testing applications will be described: the first is in the area of criterion-referenced testing and does not require IRT methods. The second application to placement testing has not previously been described in the measurement literature. Better-known adaptive testing applications in the literature include applications (1) to grading, instructional decision-making, and ability estimation (Bunderson, Inouye, & Olsen, 1989; Weiss & Kingsbury, 1984), (2) to aptitude testing (Henly & associates, 1990), and (3) to diagnostic testing (Olsen & associates, 1989).

### *Hierarchically Structured Instructional Programs*

To date, there have been only two investigations of adaptive testing to learning hierarchies that arise in objectives-based programs (Ferguson, 1969; Spineti & Hambleton, 1977). Ferguson (1969) was concerned with classifying students as "masters" or "nonmasters" on each objective in a learning hierarchy. His routing strategy was complex (involving the sequential ratio test) and required a computer to perform the actual routing. Ferguson found a 60 percent saving in number of items administered in the computerized administration using a variety of adaptive testing procedures. A test-retest of the adaptive testing procedure gave high reliability, with the reliabilities of the adaptive testing classifications higher than those of the paper-and-pencil conventional test approach.

An important consideration in the work of Spineti and Hambleton (1977) was that the adaptive testing strategies under investigation be implementable with or without the aid of computer terminals. Adaptive testing without the use of computers clearly sets this work apart from that of Ferguson and most of the other research on adaptive testing, with the exception of the self-scoring flexi-level testing work of Lord (1971). The primary effect of the restriction is that it eliminates the possibility of using

complex decision-making rules such as the one adopted by Ferguson (1969). The concern was to study the effectiveness of a multitude of adaptive testing strategies that could easily be implemented in objectives based programs. Several additional restrictions were imposed so that the results would be of maximum usefulness. A fixed number of items was required to assess mastery of each objective tested, items were scored right or wrong, and all items measuring a particular objective were assumed to have similar statistical properties. Examinee performance on the test items was assumed to be represented by the binomial test model (Lord & Novick, 1968).

The interactive effects of several factors (test length, cut-off score, and starting point) on the accuracy of mastery classification decisions and the amount of testing time in adaptive testing schemes were investigated. Values of each factor were combined to generate a multitude of adaptive testing strategies for study with two learning hierarchies and three different distributions of true scores across the hierarchies. The study was conducted via computer simulation techniques. Therefore, there was no need to be concerned about problems of developing and validating criterion-referenced tests and learning hierarchies.

Of the many learning hierarchies reported in the educational literature, two were selected for study. These were the learning hierarchies for hydrolysis of salts (Gagné, 1970) and addition-subtraction (Ferguson, 1969). The second one was selected so some of the results of this study could be compared with Ferguson's results. The two learning hierarchies are referred to as Hierarchy A and B, respectively. With Gagné's hierarchy (Hierarchy A), the adaptive testing strategies resulted, on the average, in an overall reduction of testing time of 59.2 percent. With Ferguson's hierarchy (Hierarchy B), there was a 53.2 percent reduction in testing time. It is likely that adaptive testing strategies with Hierarchy B were not quite as effective as with Hierarchy A because Hierarchy B had two terminal objectives, whereas Hierarchy A had only one. The difference highlighted the importance of the particular form of the learning hierarchy on the effectiveness of adaptive testing strategies.

The results of this study on the saving of testing time varied from 50 to 70 percent and compared favorably with the empirical results of Ferguson (1969). Ferguson reported a saving of testing time of 60 percent over conventional testing. The similarity of the results added validity to the appropriateness of the simulation procedures.

The reduction in testing time derived from the adaptive testing strategies was impressive; however, it would have meant little if the total number of errors of classification was substantially larger than with conventional testing. In fact, with Hierarchy A, the adaptive testing strategies

resulted in a slightly lower number of errors of classification than with conventional testing. The reverse was true with Hierarchy B; but, again, the differences were slight. These findings, along with the information on the comparisons of testing time for conventional and adaptive testing, provide strong support for the use of adaptive testing. That is, by using an adaptive testing strategy in the context of learning hierarchies, there is much to be gained in terms of testing efficiency without any significant loss in the accuracy of decision making.

The application of adaptive testing to learning hierarchies is substantially different from other adaptive testing applications and therefore includes some unique problems. First, there is the problem of developing and validating learning hierarchies. Because of the inter-relationship between adaptive testing schemes and a learning hierarchy, the success of any adaptive testing scheme will depend on the "validity" of the learning hierarchy under investigation. In validating learning hierarchies, there are psychological as well as statistical problems involved. For example, several researchers have reported that, while examinees may learn material in the sequence defined by a learning hierarchy, they may forget the information learned in any order. Thus, students may be able to perform a terminal objective although they have forgotten several of the prerequisite skills. The implications of this phenomenon for the validation of learning hierarchies and adaptive testing research are not clear. Second, classification problems, as opposed to measurement problems, are of interest. There has been relatively little research on using adaptive testing schemes to classify examinees into two or more categories (an exception is the work of Weiss & Kingsbury, 1984).

There are adequate technologies to develop and to validate both criterion-referenced tests and learning hierarchies (see Hambleton & Rogers, Chapter 1, and Popham, 1978). Further refinements and advancements to the technology will take place as more researchers work in the area and encounter implementation problems.

### *Basic Skills Testing*

One new promising application of adaptive testing is to the problem of placement of adults in basic educational programs in California. Currently, placement is difficult because of the following problems:

1. The very wide range of abilities represented among the examinees. This can result in a substantial amount of testing prior to effective placement.

2. The inconvenience of administering tests on a daily basis to accommodate applicants to the program. On some days there may be only one or two examinees but testing must proceed anyway.
3. Difficulties in maintaining test security.
4. General dislike of conventional tests by examinees.
5. Inefficiency in test selection, test administration, and scoring by test administrators.

All of the problems, in principle, can be overcome with computer-adaptive testing:

1. Shorten testing time dramatically by moving examinees after a 10- to 20-item routing test to test items pitched to their ability levels.
2. Permit examinees to begin testing as computer terminals are available. Testing can be initiated on an as-needed basis with minimal assistance from proctors.
3. Individualize testing (i.e., in theory each examinee sees a different test) so that test security is not a problem.
4. Provide a potentially more satisfactory test experience. At the very least, the experience of taking a test is different and so the potential for reducing test anxiety is present. Certainly, since the tests are "pitched" to ability levels, examinees should encounter less frustration.
5. Improve testing efficiency by having the computer take over the test selection, administration, scoring, and reporting functions.

The system described above is under development at the present time in the Los Angeles County School System in the United States. Preliminary results are very encouraging. Similar testing systems are currently being field-tested for college admissions tests and professional credentialing exams.

## Computer Technology

Although the literature on computer-administered testing is quite extensive, one important aspect seems to have been neglected to date. The question with which one is sooner or later confronted, when developing a computer-adaptive testing project, is the choice of hardware, peripherals, and, last but not least, the software. Yet, in the available literature one searches vainly for starting points that could make the choice between available options easier. The main purpose of this section is not to discuss the problem in detail, but instead to point out some alternative options that are available.

### *Hardware Requirements*

As a result of the state of affairs in computer technology, all early attempts to develop computer-administered testing programs were centered around large mainframe computers. These projects were typically found in the U.S. Armed Forces, large testing facilities such as Educational Testing Service and the Office of Personnel Management, and universities in the United States and in Europe. However, none of these projects lasted very long. The reasons for the failure of these early projects starting in the 1960s are manifold. First, the lack of availability of adequate peripherals such as high resolution video displays hampered the implementation of various tests, especially figural tests. A second important point was the absence of suitable operating systems that would enable the interactive use of computers with fast response times. A third point concerned the shortage of techniques for setting up an adaptive testing system. However, the main reason for abandoning the early attempts was cost benefit analyses that compared the cost of computer-administered testing programs with the traditional paper-and-pencil tests. The advances made in the 1970s and 1980s in the area of chip technology leading to a dramatic decline in the cost of hardware and the availability of cheap, small computer systems with the power of early mainframes, renewed the interest in computer-administered testing programs. The problem now is not so much the cost of hardware, but rather an ever-increasing variety of available computer systems and peripherals. In addition, one is confronted with a rapidly changing technology that makes a choice even more difficult.

Due to the state of affairs in computer technology today and the availability of cost-effective high quality peripherals, any personal computer on the market can serve as a stand-alone unit to be used in computer-adaptive testing. A basic unit would typically consist of a personal computer with a powerful 16-bit processor, a numerical co-processor, a high resolution video display and a hard disk. In addition, one would have some special peripherals such as a mouse or joy-stick or a touch-screen display. A complete system would be available for $3,000 to $5,000. Such a solution would, however, only prove adequate for small testing sites such as small commercial bureaus that would test a limited number of examinees at one time. For large testing facilities, such a system would be inadequate. Procedures for the backup of test data, administrative procedures, procedures for starting the examinee program, and supervision on examinee progress would become too cumbersome. In such a situation, another solution is needed.

The most adequate solution at this time would be to choose a system

with a main processor that can be used for development purposes, for data communication and data storage, and for supervision of examinee progress and analysis of results. Beside the main processor, the system would contain a number of independent co-processors that would each drive a test station or peripherals. The system would be designed in such a way that testing programs and necessary data can be down-loaded from the main processor to the co-processors that can then operate independently. Such systems are available from a number of vendors. The advantages of such an architecture are obvious. One combines the flexibility of stand-alone systems while avoiding the problems of communication between stand-alone systems. In addition, the system is easily expandable with new co-processors when needed.

One of the problems that still exists today in spite of the advances in chip technology and the quality of peripherals, especially video displays, is the presentation of figural or graphic items. Usually, video displays use what is called bit-mapped graphics. This means that a display consists of a number of points that are on or off, thus creating an image on the screen. A display with a moderate resolution would consist of a bit map of 600 by 400 points, while a high resolution display would have a resolution of 1,024 by 1,024 points. One item would therefore consist of 240,000 to 1,000,000 bits of information. If one uses a color display, these figures should be multiplied by a factor of 3 to 8, depending on the number of different colors used. A typical test consisting of 200 figural items, not uncommon in an adaptive testing program, would take between 6Mb and 30Mb of computer memory. Despite the advances in the area of chip technology, no personal computer available today on the market would have such an amount of memory on line. Therefore, the information from items would have to be loaded in memory, say, from a hard disk, and thus slow down the program, especially if several examinees were taking the same tests.

Alternatives used today are, for example, the use of video devices. A problem when using video devices such as cassettes or tapes arises, however, if one wishes to mix video images with computer-generated text. This generates a complicated synchronization problem that has yet to be solved in an elegant way. Therefore, one should be cautious to use video images if these images have to be combined with computer-generated text. An elegant solution for both problems is the use of so-called laser vision technology. Graphic images recorded on videotape can be reproduced on a laser vision disk that is computer controlled. The advantages of this system are: high quality display, large storage facilities (10,000 images per disk), fast display rates, and a high degree of flexibility. Additional features of a video disk are that, apart from displaying one frame at a time, the com-

puter can instruct the video player to show a number of frames in sequence, thus enlarging the scope of the present tests to moving images.

A system using a multiprocessor architecture combined with laser vision technology is presently in development at the Rijks Psychologische Dienst in the Netherlands.

### *Software Requirements*

Discussions regarding software requirements of computer-adaptive testing projects usually center around issues like exchangeability of software and the choice of a specific language. However, as ready-to-use packages are still not widely available, the problem regarding the software should, in our view, be dealt with in a pragmatic manner. The choice of a specific language such as PASCAL, BASIC, FORTRAN, FORTH, or any other language available in microcomputers today is not the main problem as long as the development tools available are adequate and the compiler used produces efficient and fast execution programs.

The main problem is to develop a flexible system that is not only adequate for present applications but is also designed in such a way that future developments in the area of ability testing can be easily implemented or at least not inhibited by the original design. In addition, a system should include all components necessary in the process between planning the test and the report regarding the results of the examinees' performance. In short, the system should include the following components: administrative procedures; item banking; test construction; test presentation; test scoring; and report formatting.

In the administrative section, data is recorded such as test date and examinees' data such as name, address, which tests are to be given and in what sequence and in what format. The item banking module should enable the user to add, update, and change items in an easy and flexible manner including item text and item parameters.

The test construction module is used to compile tests and test batteries. In addition, this module also records additional data regarding the execution of the test such as test format (adaptive testing, speeded, fixed sequence of items, recording of response times, norms, etc.). The test presentation module is the core of the program. It controls the presentation of items and tests in the specified order and records the examinees' answers, response times, and so on. In an adaptive format, it controls the sequence of items

according to a specified rule or algorithm. The test scoring module transforms responses to test scores and normative scores, and prints out all results in a test profile for the examinee and/or test user. The test results may be interpreted into a more or less extensive report.

All modules should be part of an integrated program that should operate in a user friendly manner, preferably menu driven. The actual code for running a test station including the necessary data should be automatically produced by the program and be downloaded in the co-processor. The program should be constructed in such a way that one can follow the program flow in the co-processors from the main processor. Ideally, the system should be constructed in such a way that the program can be stopped and restarted from any point.

Computer-administered testing programs that are more or less closely constructed according to these guidelines are in development or operation in the U.S. Navy, the Belgian Police Selection Institute, the German Army Forces, and the Dutch Rijks Psychologische Dienst. The latter service is developing a system that resembles these guidelines and the previously described hardware most closely. At this time, it is too early to consider some sort of standardization or exchangeability. First, the systems that are developed or are already in operation should prove their merit, as more experience is gathered in operating computer-assisted programs. A definite choice between one of the available options today would be extremely imprudent merely because of the rapidly changing computer technology, especially in the area of graphic and video processors.

A number of computer packages are now available to support computerized adaptive testing (for a review, see Hsu & Yu, 1989). Probably the most comprehensive package at the present time is MicroCAT (Assessment Systems Corporation, 1988). According to Stone (1989), an integrated testing system should support item and test development, test administration, item and test analysis, and reporting test results. MicroCAT (Version 3.0) is an integrated test development, administration, and data analysis system that addresses all of Stone's concerns, as well as our earlier stated requirements for a desirable computer-assisted testing system. MicroCAT runs on microcomputers, is menu-driven, and has been in use since about 1980. Probably because of its 10-year use and up-dating, the current system contains many features that users desire and is free of errors that often plague newer systems. Features include (1) handling of multiple-item formats and items with graphics, (2) basing test development and analyses on either classical test theory or IRT principles and methods, and (3) constructing, administering, and scoring fixed length or adaptive tests.

## Guidelines for Evaluating Computer-Adaptive Testing

Two main questions about computer-adaptive testing that arise concern (1) the relationship between scores and associated decisions with conventionally administered and computer-administered tests, and (2) the viability of IRT models for providing a technically adequate measurement system. In the spirit of insuring that computer-adaptive tests perform as well as paper-and-pencil tests for an important military testing program in the United States, Green, Bock, Humphreys, Linn, and Reckase in 1982 provided a set of guidelines for evaluating computer-adaptive tests. These guidelines, however, are generally applicable to all forms of computer-administered tests. The authors divided the guidelines into nine main categories: content considerations, dimensionality, reliability, validity, item parameter estimation and linking, item pool characteristics, item selection and test scoring, and human factors. A selected list of the guidelines, organized by category from Green and associates (1982, 1984), is contained in Figure 12–1. Also, APA (1986) prepared a set of guidelines which apply more generally to computer-administered tests.

Recently, both Green (1988) and Wainer and Kiely (1987) have highlighted a number of difficult problems to overcome in CAT systems. Green was especially concerned with problems in the areas of equating and item selection. The equating problem arises because, while two tests, a paper-and-pencil test and an adaptive test, can be equated or calibrated on the same scale using an IRT model, in general, the two tests will not provide the same degree of measurement precision at points along the ability continuum (i.e., the information functions for the two tests will, in general, differ). Therefore, the students being assessed will not (or should not) be indifferent as to which test they prefer. Good students should prefer the test leading to the most accurate scores; poor students should not. The implication of this non-equivalence of the measurement properties of a paper-and-pencil test and an adaptive test remains a problem in practice.

The problem in item selection arises because of a concern that the context in which an item appears (e.g., item position) and its content may influence item performance. If, for example, item performance is influenced by items that may have been administered previously, then valid comparisons of examinees, when examinees are not administered the same items—and, in general, they will not be in adaptive testing—are problematic. The influences of item context and content represent threats to the validity of IRT models and, specifically, the invariance of both item and ability parameter estimates. Wainer and Kiely (1987) also documented the context effect of (1) item parameter estimates due to an item's location in

Figure 12–1. Guidelines for Evaluating Computer-Adaptive Testing (CAT). These guidelines were prepared by Green and associates (1982). A sample of their guidelines is offered here, with minor editing to enhance their readability without knowledge of the full text of their report.

---

*Content Considerations*

1. Specifications for item content should be the same for CAT and paper-and-pencil tests.
2. The content of items selected for the item pool should match the content specifications.
3. Test items must be designed to match the capabilities of the computer equipment.

*Dimensionality*

4. The fit of the IRT model should be checked.
5. Highly discriminating items should be selected.
6. A factor analysis of the inter-item tetrachoric correlations should be performed.
7. Local independence assumption should be examined.
8. Subtests should be formed when tests are not unidimensional.
9. A test should be balanced to reflect the heterogeneity of domain content and item formats.

*Reliability*

10. The standard error of measurement of each test score should be reported as a function of the test score, in the metric of the reported score.
11. The standard error of measurement of each test should also be reported in the ability metric.

*Validity*

12. The similarity of variance-covariance matrices for CAT and paper-and-pencil tests should be assessed.
13. The covariance-structures of the two versions should be compared.
14. The CAT and paper-and-pencil versions of a test should be validated against the same external criteria.
15. The extent of prediction bias should be assessed for important subpopulations.

*Item Parameters—Estimation*

16. The sample for item calibration should be of adequate size, currently at least 1,000 cases.
17. The calibration sample should be selected so that a sufficient number of examinees are available in the range of ability needed to estimate the lower asymptote and the point of inflection of the item characteristic curve.

Fig. 12–1. Continued

18. The procedure for estimating item parameters should be shown to be "empirically consistent" (large samples should lead to good estimates).
19. The procedure for estimating item parameters should be shown to be unbiased, or the extent and nature of the bias should be specific.
20. The item characteristic curves should fit the observed data.
21. The difficulty of items administered in the CAT and paper-and-pencil versions should be compared.

*Item Parameters—Linking*

22. The linking procedure for placing items on a common scale should be fully described.
23. When using an equivalent groups procedure for linking, the equivalence of groups should be demonstrated.

*Item Pool Characteristics*

24. The distribution of the item parameter estimates and descriptive statistics for the estimates should be presented.
25. The information for the total item pool should be presented.

*Item Selection and Test Scoring*

26. The procedure for item selection and ability estimation must be documented explicitly and in detail.
27. The procedure should include a method of varying the items selected, to avoid using a few items exclusively.
28. The computer algorithm must be capable of administering designated items, and recording the responses separately, without interfering with the adaptive process.
29. The computer must be able to base the choice of a first item on prior information.

*Human Factors*

30. The environment of the testing terminal should be quiet and comfortable, and free of distractions.
31. The display screen should be placed so that it is free of glare.
32. The legibility of the display should be assessed empirically.
33. The display must be able to include diagrams that have fine detail.

a test, and (2) ability estimates due to the sequence and emphasis (or de-emphasis) of specific test content.

Wainer and Kiely (1987) proposed that "testlets," rather than items, become the building blocks for CAT based on multistage fixed branching. Here, a testlet is

> . . . 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. In this way, each item is embedded in a predeveloped testlet, in effect carrying its own context with it. The paths through a testlet may follow a hierarchical branching scheme that routes examinees to successive items of greater or lesser difficulty depending on their previous responses and culminates in series of ordered score categories. Or the testlet may contain only a single linear path of a number of items that are administered to all examinees. The form chosen depends critically on the application for which it is intended.
>
> Just as branching schemes may vary within testlets, the testlets themselves may also be combined to form a complete test by linking them hierarchically or in a linear fashion, or some combination of the two, again depending on the intended purpose of the test. This arrangement allows for the construction of a wide variety of tests for specific purposes by combining hierarchical and linear branching both between and within testlets in any desired combination (Wainer and Kiely, 1987, pp. 190–191).

Clearly, considerably more research is needed to both document the strengths and shortcomings of IRT-based CAT models as well as to pursue alternate CAT models such as testlets within an IRT framework.

## Conclusions

The promise of adaptive testing has been high and now there is substantial empirical evidence to suggest that the expected advantages are being obtained in practice. In addition, Green and his associates (1982, 1984) have provided a set of guidelines for insuring that computer adaptive tests function as well as their paper-and-pencil counterparts. Finally., there exists quality computer software such as MicroCAT to support computer adaptive testing. For all of these reasons, computer adaptive testing has arrived and expanded use can be expected.

A number of topics for further research seem especially appropriate at this time. Readers are referred to Wainer (1990, Chapter 9) for a comprehensive list. Only three topics will be presented here. First, the topic of "content matching" between computer adaptive tests and their paper-and-pencil counterparts needs to be better understood. Most of the developmental work to date has not worried much about this problem, especially when the measurements from the two tests were comparable. But full acceptance of adaptive tests by examinees and test users, especially of very important tests, may require that adaptive tests adhere to the same content specifications as their full length paper-and-pencil counterparts.

A second important problem for study concerns the modelling of student item performance (Green, 1988). Successful adaptive testing based on IRT principles and methods requires an accurate match between the psychometric model used to account for examinee performance and actual examinee performance. Unidimensional models with more than three item parameters (de Ayala, 1989) or even multidimensional models (Reckase, 1989) may prove to be more valuable than one-, two-, and three-parameter logistic models which are in current use. More model-building research seems highly desirable at this time.

Finally, additional research results with respect to several of the important components of adaptive testing such as starting places, item selection, ability estimation, and stopping rules would seem to be in order. There are still many questions about the implications of these various components on the success of adaptive testing.

## References

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

Assessment Systems Corporation. (1988). *User's manual for the MicroCAT Testing System, Version 3.* St. Paul, MN: Author.

Berk, R. A. (Ed.). (1984) *A guide to criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Brzezinski, E., & Hiscox, M. (Eds.). (1984). Microcomputers in educational measurement (Special Issue). *Educational Measurement: Issues and Practice 3*:3–50.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (ed.), *Educational measurement*, 3rd ed. New York: Macmillan, pp. 367–407.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement 49*:789–805.

Ferguson, R. L. (1969). The development of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh.

Gagné, R. M. (1970). *The conditions of learning*, 2nd ed. New York: Holt, Rinehart, & Winston.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (eds.), *Principles of modern psychological measurement: A Festschrift for Frederic M. Lord.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 69–80.

———— (1988). Critical problems in computer-based psychological measurement. *Applied Measurement in Education 1*:223–231.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D.

Hambleton, R. K. (ed.). (1983). *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (ed.), *Educational measurement*, 3rd ed. New York: Macmillan, pp. 147–200.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer Academic Publishers.

Henly, S. J., Klebe, K. J., McBride, J. R., & Cudeck, R. (1990). Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement 13*:363–371.

Hsu, T. C., & Yu, L. (1989). Using computers to analyze item response data. *Educational Measurement: Issues and Practice 8*:21–27.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education 2*(4):359–375.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement 8*:147–151.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement 21*:315–330.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery and computerized adaptive testing subtests. *Applied Psychological Measurement 8*:155–163.

Olsen, J. B., Cox, A., Price, C., & Strozeski, M. (1989, April). *Development, implementation, and validation of a predictive and prescriptive test for statewide assessment*. Paper presented at the meeting of AERA, San Francisco.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association 70*:351–356.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

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

Spineti, J. P., & Hambleton, R. K. (1977). A computer simulation study of tailored testing strategies for objectives-based instructional programs. *Educational and Psychological Measurement 37*:139–158.

Stone, C. A. (1989). Testing software review: MicroCAT Version 3.0. *Educational Measurement: Issues and Practice 8*:33–38.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement 14*:181–196.

———— (1983, August). *Tailored testing theory and practice: A basic model, normal ogive submodels, and tailored testing algorithms* (NPRDC TR 83-82).

San Diego, CA: Navy Personnel Research and Development Center.

Wainer, H. (1988). The first four millennia of mental testing: From ancient China to the computer age. In *Proceedings of 29th Annual Military Testing Association Conference*, pp. 357–362.

———— (ed.). (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement 24*:185–201.

Ward, W. C., Kline, R. G., & Flaugher, J. (1986). *College Board Computerized Placement Tests: Validation of an adaptive test of basic skills* (Research Report 86-29). Princeton, NJ: Educational Testing Service.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement 6*:473–492.

———— (ed.). (1983). *New horizons in testing.* New York: Academic Press.

———— (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology 53*:774–789.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement 21:* 361–375.

WICAT Systems (1986). *Educational measurement system.* Orem, UT: Author.

(1982, May). *Evaluation plan for the Computerized Adaptive Vocational Aptitude Battery.* Baltimore, MD: The Johns Hopkins University, Department of Psychology.

———— (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement 21*(4):347–360.