Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education

David J. Weiss
University of Minnesota

Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education

David J. Weiss

Computerized adaptive testing (CAT) is described and compared with conventional tests, and its advantages summarized. Some item response theory concepts used in CAT are summarized and illustrated. The author describes the potential usefulness of CAT in counseling and education and reviews some current issues in the implementation of CAT.

Effective counseling in education frequently requires that detailed information on a variety of traits and characteristics of each student be available. In the early days of counseling, when counseling revolved around vocational issues (e.g., Brayfield, 1961; Lofquist & Dawis, 1969), student information was primarily concerned with vocationally relevant abilities and preferences. As counseling expanded beyond the vocational realm and education expanded its focus beyond just academic skills, the range of information used by counselors and educators included all aspects of personality, needs, values, attitudes, and, most recently, interpersonal and relationship variables.

In some counseling and educational situations, some of this information can be gleaned by a skilled counselor from data obtained through interviews or by a teacher from observations of students' behavior. However, in many cases, counselors and educators rely on a wide variety of psychological measuring instruments—tests, questionnaires, inventories, and scales—to provide measurement data to inform counseling and educational services. Research on counseling—as well as research in education and development—relies heavily on these measuring instruments for its data.

CONVENTIONAL PSYCHOLOGICAL MEASURING INSTRUMENTS

The vast majority of psychological measurement instruments in use today are based on the paper-and-pencil conventional test. This type of test was developed initially for use in World War I to provide a quick and inexpensive method of screening large numbers of recruits (DuBois, 1970). Tests of this type (including most inventories and other instruments used to measure nonability variables) are typically designed using procedures of classical test theory (e.g., Cronbach, 1990; Gulliksen, 1950), which has its roots in procedures that developed around the same time as the paper-and-pencil test.

A conventional test is characterized by a fixed set of questions or items that are administered to each examinee. Using the guidance of classical test construction procedures, the developers of conventional tests usually select these items by item analysis procedures that are designed to maximize the internal consistency reliability of the set of items that make up the instrument. Although the reliability will be high for that set of

David J. Weiss, Department of Psychology, University of Minnesota. For further information on adaptive testing, visit www.psych.umn.edu/psylabs/CATCentral/. Correspondence concerning this article should be addressed to David J. Weiss, N660 Elliott Hall, University of Minnesota, Minneapolis, MN 55455-0344 (e-mail: djweiss@umn.edu).

items, which also reduces the standard error of measurement, this reduction in measurement imprecision is assumed to be constant across the measurement scale.

There are, however, a number of important limitations of conventional tests. When classical test theory is used to select the items for a conventional test, the items selected that maximize internal consistency reliability are typically those that are appropriate for the average examinee in the group—the items that maximize reliability are those that have their difficulties (proportion correct or keyed) around $p = .50$. These items are those that provide best measurement for the average examinee, but they are too difficult for examinees who are below average on the trait being measured and are too easy for examinees who are above average.

For example, if a test is designed to measure arithmetic ability and it is designed for fourth graders, it will be too difficult for most second graders and too easy for most sixth graders. Yet, in the fourth-grade class, there are likely some students who are functioning at or below the second-grade level and some functioning at or well above the sixth-grade level. For the students who deviate in ability/achievement from the level of the conventional test, the test will provide very little information. The students with low ability will answer almost all of the items incorrectly, and the students with high ability will answer all or most of the items correctly. The result, for these students, will be scores that provide almost no capability of differentiating among them.

This same principle applies in the measurement of all psychological variables that can be measured as continuous variables—a fixed-item conventional measuring instrument is designed to measure well for a restricted range of the trait, usually around the mean of the anticipated trait distribution. When it is used for individuals whose trait levels deviate from that trait range, conventional measuring instruments provide increasingly poor measurement because the items have little relevance for those examinees; this has been recognized in the application of "out-of-level" testing in some educational environments. Furthermore, time limits that are frequently imposed on conventional tests (usually for the test administrator's convenience) further deteriorate the quality of measurement by introducing other traits (e.g., persistence, slowness) that interfere with good measurement of the trait(s) that the instrument is designed to measure.

ADAPTIVE TESTING

The basic measurement problem that characterizes conventional tests has been recognized for many years in a number of domains. In athletic competitions, for example, it would be unheard of to try to measure an athlete's hurdle-jumping ability by having her or him repeatedly jump over a succession of 2-foot hurdles. Rather, a series of hurdles of increasingly higher levels is set up, and the athlete tries to clear each until she or he is no longer able to do so. Then, to determine a more precise indication of the level that the participant can clear, a set of hurdles that vary in a relatively narrow range around the level at which the individual began to miss is constructed. In this way, the task is "adapted" to the individual's performance in order to obtain precise estimates of each athlete's ability.

This principle of adapting the test to the examinee was recognized in the very early days of psychological measurement, even before the development of the conventional paper-and-pencil test, by Alfred Binet in the development of the Binet IQ test (Binet & Simon, 1905) that later was published as the Stanford-Binet IQ Test. Binet's test comprised sets of test items normed by chronological age level. He selected items for each age level if approximately 50% of the children at that age level answered an item correctly. Thus, in the original version of the test, there were sets of items at ages 3 years through 11 years. All of these items constituted Binet's item "bank" for his adaptive test.

Binet's test administration procedure is a fully adaptive procedure:

1. It uses a precalibrated bank of test items.
2. It is individually administered by a trained psychologist and is designed to "probe"

for the level of difficulty (i.e., chronological age) that is most appropriate for each examinee, much as jumping hurdles probes for the performance level of each athlete.

3. It has a variable starting option. The administrator sets the beginning level of the Binet test on the basis of her or his best guess about the examinee's likely level of ability (typically the examinee's chronological age, but the starting level can be lower or higher if there is information to inform such a starting level).
4. It uses a defined scoring method—a set of items at a given age level is administered and immediately scored by the administrator.
5. There is a "branching," or item selection rule, that determines which items will next be administered to a given examinee. In the Binet test, the next set of items to be administered is based on the examinee's performance on each previous set of items. If the examinee has answered some or most of the items at a given age level correctly, usually the items at the next higher age level are administered. If most of the items at a given age level are answered incorrectly, items at the next lower age level are typically administered.
6. There is a predefined termination rule. The Binet test is terminated when, for each examinee, both a "ceiling" and a "basal" level have been identified. The ceiling level is the age level at which the examinee incorrectly answers all items; the basal level is the age level at which the examinee correctly answers all the items. The effective range of measurement for each examinee lies between these two levels.

An examinee's final score on the Binet test is based on the subset of items that she or he answered correctly. In effect, these items are weighted by their age level in arriving at the IQ scores derived from the test, because different examinees will answer both different numbers and subsets of items.

COMPUTERIZED ADAPTIVE TESTING (CAT)

Binet's adaptive approach to ability measurement remained the only operational adaptive test for more than a half century because the requirements of a world war resulted in the conventional paper-and-pencil technology that dominated psychological and educational testing for most of the 1900s. In the 1950s, however, U.S. Army researchers began exploring the possibility of delivering rudimentary adaptive tests using both paper and pencil and testing machines (Bayroff, 1964; Bayroff, Thomas, & Anderson, 1960). Both of these approaches were unsuccessful, and adaptive testing survived only in the Binet tests.

In the late 1960s, the Personnel and Training Research Programs of the U.S. Office of Naval Research began supporting theoretical research on item response theory (IRT) and adaptive testing by Frederic Lord (e.g., Lord, 1970, 1971a). This was paralleled by some early "tailored" testing research, both in the context of computer-assisted instruction (e.g., Ferguson, 1969) and in measuring educational achievement (e.g., Cleary, Linn, & Rock, 1968, 1969), and in an applied CAT research program also supported by the U.S. Office of Naval Research (Weiss & Betz, 1973), resulting in the beginning of the now burgeoning field of CAT.

CAT (Van der Linden & Glas, 2000; Wainer et al., 2000) is the redesign of psychological and educational tests for effective and efficient administration by interactive computers. Its objective is to select, for each examinee, the set of test questions that simultaneously most effectively and efficiently measure that person on the trait. CAT builds on and improves upon Binet's implementation of adaptive testing by replacing the human administrator with a computer program. Test items are stored in the computer and displayed on its monitor, and the examinee interacts with the test either by keyboard or mouse. The computer program functions much like the psychologist administering a Binet test—it determines how to begin the test for a given examinee, selects items based on the examinee's scored responses to previous items, and applies one or more rules to terminate an examinee's test. Early approaches to CAT were based on variations of Binet's approach (Weiss, 1973) or other "branched" approaches that are based on different ways of structuring an item bank (e.g., Lord, 1971b, 1971c). It soon became apparent that each of these approaches had its problems (Weiss, 1974) and that these problems could be resolved by methods of IRT, which was maturing during the 1970s.

Some IRT Concepts Used in CAT

IRT is a family of mathematical models that describe how people interact with test items (Embretson & Reise, 2000). These models were originally developed for test items that are scored dichotomously (correct or incorrect), but the concepts and methods of IRT extend to a wide variety of polytomous models for all types of psychological variables that are measured by rating scales of various kinds (Van der Linden & Hambleton, 1997).

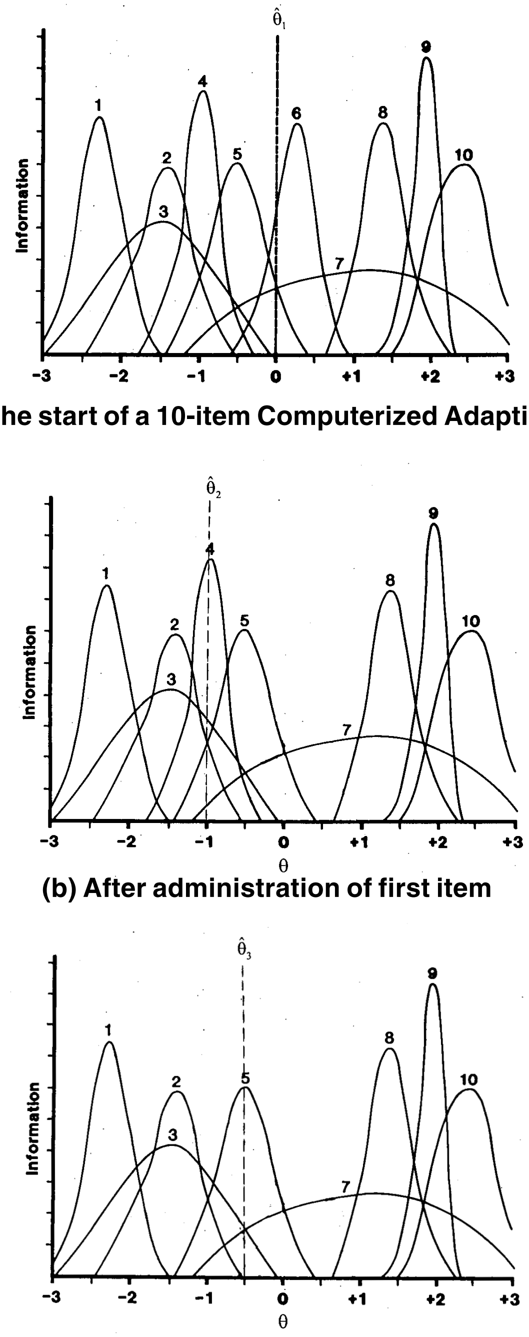
In the context of items scored correct-incorrect, test items are described by their characteristics of difficulty and discrimination, as they are in traditional item and test analysis. However, in IRT, these item statistics (referred to as "parameters") are estimated and interpreted differently than classical proportion correct and item-total correlation. For multiple-choice items, IRT adds a third item parameter referred to as a "pseudo-guessing" parameter that reflects the probability that an examinee with a very low trait level will correctly answer an item solely by guessing.

Information. Although these three item parameters are useful in their own right, for purposes of CAT they are combined into an "item information function" (IIF). The IIF is computed from the item parameters. It describes how well, or precisely, an item measures at each level of the trait that is being measured by a given test (referred to in IRT by the Greek letter theta, θ). A major advantage of IRT is that both items and people are placed on the same scale (usually a standard score scale, with $M = 0.0$, and $SD = 1.0$) so that people can be compared to items, and vice versa, by determining the distance of a person's trait level to each item's location on the same continuum.

Figure 1, part a, shows IIFs for 10 items. The location of the center of the IIF reflects the difficulty of the item, the height of the IIF reflects the item discrimination, and its asymmetry reflects the magnitude of the pseudo guessing parameter. Thus, because Item 1 is the easiest item, its IIF is on the left end of the θ continuum, and Item 10 is the most difficult. Because Item 9 is the most discriminating, it has the highest IIF, and Item 7 is the least discriminating. None of these items have a high pseudoguessing parameter because all of the IIFs are reasonably symmetric.

IRT scoring. Whereas tests not using IRT are typically scored by counting the number of correct answers provided by an examinee, or if using rating scale items by summing a set of arbitrary weights assigned to the response categories of each item, IRT uses a quite different method of scoring or estimating θ levels for examinees. This method is called "maximum likelihood estimation" (MLE). In contrast to number-correct scoring, MLE weights each item by all three of its item parameters and also considers whether the examinee answered each item correctly. In some cases, it might be appropriate to incorporate into θ estimation prior information about an examinee or an assumed prior distribution of θ. When such prior information is included in θ estimation, along with the examinee's scored responses and the characteristics of each item, MLE becomes a Bayesian procedure, such as EAP (expected a posteriori) or MAP (maximum a posteriori) estimation (see Embretson & Reise, 2000, chap. 7, for further explanation of these scoring procedures).

As a result of combining information on the examinee's entire pattern of responses as well as the characteristics of each item, MLE can provide many more distinctions among examinees than can number-correct scoring. For example, number-correct scoring of a 10-item conventional test can result in at most 11 scores (0 to 10); MLE for the same test can result in 2^{10} (or 1,024) different θ estimates ($\hat{\theta}$).



(a) At the start of a 10-item Computerized Adaptive Test

(b) After administration of first item

(c) After administration of two items

FIGURE 1

Item Information Functions and θ Estimates

MLE has an additional advantage over number-correct scoring. In addition to providing a $\hat{\theta}$ for each examinee, MLE also provides an individualized standard error of measurement (SEM) for each $\hat{\theta}$. Unlike the SEMs from non-IRT test analysis methods, the SEMs from IRT can vary from person to person, depending on how she or he answered a particular set of items. Finally, the $\hat{\theta}$s and their SEMs in IRT are not dependent on a particular set of items—they can be determined from any subset of items that an examinee has taken, as long as the parameters for those items have been estimated on the same scale.

IRT-Based CAT

In IRT-based CAT, a relatively large item bank is developed for a given trait and the items' IIFs are determined. Similar to Binet's item bank, a good CAT item bank has items that collectively provide information across the full range of θ. An examinee begins the CAT with an initial θ estimate, which can be the same for all examinees or like the Binet procedure can use any prior information available on the examinee. Also, as with the Binet test, an item is administered and immediately scored, but by the computer that is delivering the test rather than by a human test administrator.

Item selection. At this point in the process, IRT-based CAT deviates from the Binet approach. Rather than administering a set of items before "branching" to a different set of items, CAT selects each next item based on the examinee's scored responses to all previous items. At the initial stages of a CAT, when only a single item or two have been administered, the next item is usually selected by a "step" rule—if the first item was answered correctly, the examinee's original prior $\hat{\theta}$ is increased by some amount (e.g., .50); if the first item was answered incorrectly, the original $\hat{\theta}$ is decreased by the same amount. As the test proceeds and the examinee obtains a response pattern of at least one correct and one incorrect response, MLE is used to obtain a new $\hat{\theta}$, which is estimated from the examinee's responses to all administered items at that point in the test.

After each item is administered and scored, the new $\hat{\theta}$ is used to select the next item. That item is selected from all items in the item bank that have not previously been administered to that examinee by identifying the item that provides the most information at the current $\hat{\theta}$. Figure 1 illustrates "maximum information" item selection in CAT. In addition to displaying information functions for 10 items, Figure 1, part a, shows an initial $\hat{\theta}_1$ for a hypothetical examinee. This value is shown at 0.0, which is the mean of the θ scale. Values of information are computed for all items at that θ level. Figure 1, part a, shows that Item 6 provides the most information of the 10 items at $\theta = 0.0$. Therefore, Item 6 is administered and scored. On the basis of that score (incorrect, in this example), a new $\hat{\theta}_2$ is determined (in this case using a step size of 1.0) as –1.0. On the basis of the maximum information item selection rule, Item 4 is administered (Figure 1, part b) and scored. Assuming that Item 4 was answered correctly, MLE can now be used to estimate θ. The result is $\hat{\theta}_3 = -.50$. Again, selecting an item by maximum information results in the selection of Item 5 (Figure 1, part c). Scoring, θ estimation, and item selection continue (like the Binet test) until a termination criterion is reached.

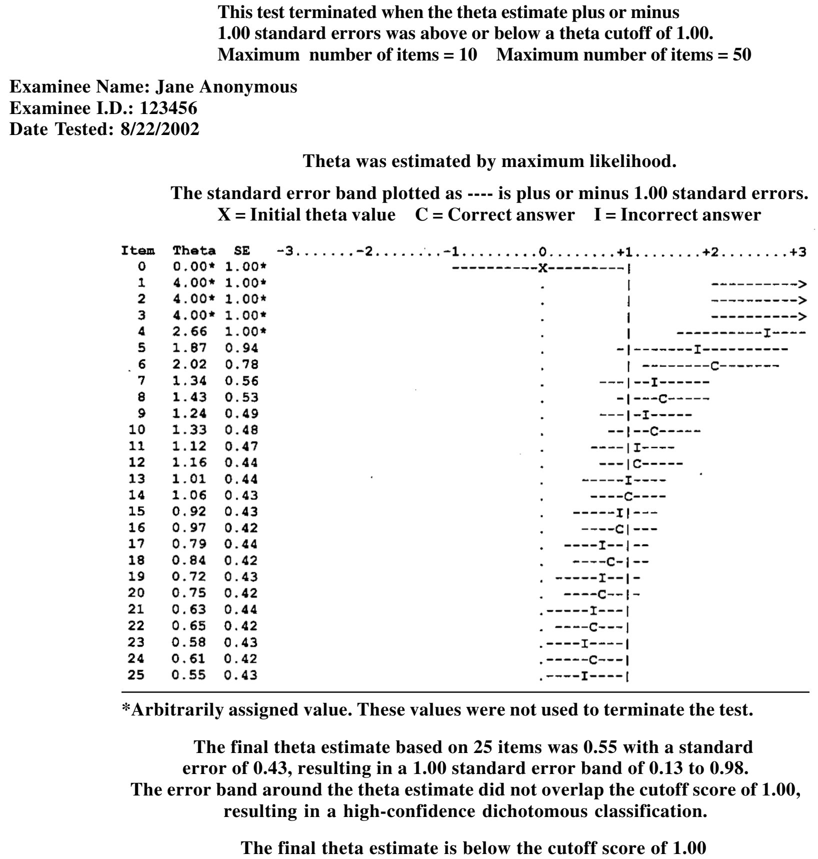
Ending a CAT. One important characteristic of CAT is that the test termination criterion can be varied for different testing objectives. Some tests are used for selection or classification, for example to classify an individual as having mastered some domain of achievement or to select individuals who will be admitted to a school or college. Other tests are used for counseling or clinical purposes. The objective of such tests is to measure each individual as well as possible. In the context of CAT, these two objectives are operationalized by two different termination rules.

For classification purposes, an individual's score is compared against some cutoff value. The objective is to make a classification that is as accurate as possible, for example, to classify examinees as "masters" or "nonmasters" of an achievement domain with no more than a 5% error rate. To implement this in the context of CAT, both the $\hat{\theta}$ and its associated SEM are used. An individual can be classified as being above a cutoff value (expressed on the θ

scale) if both the $\hat{\theta}$ and its 95% confidence interval (calculated as plus or minus two SEMs) are above or below the cut score. Because CAT can evaluate this decision after each item is administered, the test can be terminated when this condition is satisfied. The result of such a test will be a set of classifications made for a group of examinees that all have at most a 5% error rate. The error rate can be controlled by changing the size of the SEM confidence interval around $\hat{\theta}$.

When CATs are not used for classification, a different termination rule applies. In this case, it is desirable to measure each examinee to a desired level of precision, as determined by a predetermined level of SEM. This will result in a set of "equiprecise" measurements, such that all examinees will have scores that are equivalently accurate—perhaps defining a new concept of "test fairness." To implement equiprecise measurement, CAT allows the user to specify the level of SEM that is desired for each examinee. Assuming that the item bank has a sufficient number of test items properly distributed across the θ scale and that the test is allowed to continue long enough for the examinee, this goal will be achieved if the test is terminated when that level of SEM is reached.

A sample CAT. Figure 2 shows the response record of a single examinee's progress through a CAT (Assessment Systems Corporation, 2001). This CAT was designed to make a dichotomous classification around θ = 1.0 (1 standard deviation above the mean), with a ±1 SEM



This test terminated when the theta estimate plus or minus
1.00 standard errors was above or below a theta cutoff of 1.00.
Maximum number of items = 10   Maximum number of items = 50

Examinee Name: Jane Anonymous
Examinee I.D.: 123456
Date Tested: 8/22/2002

Theta was estimated by maximum likelihood.

The standard error band plotted as ---- is plus or minus 1.00 standard errors.
X = Initial theta value   C = Correct answer   I = Incorrect answer

```
Item  Theta  SE   -3.......-2........-1.........0........+1........+2........+3
  0   0.00* 1.00*                     ----------X---------|
  1   4.00* 1.00*                               .         |         ---------->
  2   4.00* 1.00*                               .         |         ---------->
  3   4.00* 1.00*                               .         |         ---------->
  4   2.66  1.00*                               .         |     ----------I----
  5   1.87  0.94                                .        -|-------I----------
  6   2.02  0.78                                .         | --------C-------
  7   1.34  0.56                                .      ---|--I------
  8   1.43  0.53                                .        -|---C-----
  9   1.24  0.49                                .      ---|-I-----
 10   1.33  0.48                                .       --|--C-----
 11   1.12  0.47                                .     ----|I----
 12   1.16  0.44                                .      ---|C-----
 13   1.01  0.44                                .    -----I----
 14   1.06  0.43                                .     ----C----
 15   0.92  0.43                                .   -----I|---
 16   0.97  0.42                                .    ----C|---
 17   0.79  0.44                                .  ----I--|--
 18   0.84  0.42                                .   ----C-|--
 19   0.72  0.43                                . -----I--|-
 20   0.75  0.42                                .  ----C--|-
 21   0.63  0.44                                .-----I---|
 22   0.65  0.42                                . ----C---|
 23   0.58  0.43                                .----I----|
 24   0.61  0.42                                .-----C---|
 25   0.55  0.43                                .----I----|
```

***Arbitrarily assigned value. These values were not used to terminate the test.**

The final theta estimate based on 25 items was 0.55 with a standard
error of 0.43, resulting in a 1.00 standard error band of 0.13 to 0.98.
The error band around the theta estimate did not overlap the cutoff score of 1.00,
resulting in a high-confidence dichotomous classification.

The final theta estimate is below the cutoff score of 1.00

FIGURE 2

Item-by-Item Report of Maximum Information Computerized Adaptive Testing for a Single Examinee

band (a 68% confidence interval). The initial θ estimate (X) was 0.0, and the test item providing maximum information at that θ level was administered and answered correctly (C). The initial step size was 3.0 to attempt to force a mixed (correct-incorrect) response pattern as quickly as possible, so the next item had maximum information at $\theta = 3.0$. It, too, was answered correctly, so additional difficult questions were given (Items 3 and 4) until an incorrect answer (I) was obtained. At that point, MLE was used to obtain a $\hat{\theta}$ of 2.66. The item at that level (Item 5) was also answered incorrectly, and the resulting $\hat{\theta}$ was 1.87 with an SEM of .94. Note that each time a correct answer was obtained, $\hat{\theta}$ increased and that an incorrect answer led to a decrease in $\hat{\theta}$. Note also that the differences between successive $\hat{\theta}$s decreased as the test proceeded, indicating that the test was converging on the examinee's θ level; also, in general, the SEM tended to decrease, because additional item responses generally improve the estimation of θ.

In this test, the examinee's $\hat{\theta}$ followed a downward trend, falling below the cut score of $\theta = 1.0$ at Item 15. However, the $\hat{\theta}$ could not be assumed to be reliably below that cut score because the SEM band still included $\theta = 1.0$ and so the test continued for another 10 items until the examinee's $\hat{\theta}$ and its SEM were entirely below the cut score. This occurred at Item 20 (.55 + .43 = .98, which is just below 1.00), and the test was terminated. The test results indicated that this examinee's $\hat{\theta}$ was below the cut score, with at least 68% confidence (actually, in this case because that confidence interval was symmetric and 50% was below the mean, the confidence level of a unidirectional decision would be 50% + 34% = 84%). Higher confidence could have been obtained by using a 2 SEM interval around $\hat{\theta}$, which obviously would have a required a longer test.

Figure 2 also illustrates another characteristic of most CATs: As the test progresses, the examinee tends to alternate between correct and incorrect answers, as can be seen beginning with Item 7 or 8. This is the result of the convergence process that underlies CAT. The result, typically, is that each examinee will answer a set of questions on which he or she obtains approximately 50% correct, even though each examinee will likely receive a set of questions of differing difficulty. In a sense, this characteristic of a CAT tends to equalize the "psychological environment" of the test across examinees of different trait levels. By contrast, in a conventional test, the examinee who is high on the trait will answer most items correctly and the examinee who is low on the trait will answer most of the items incorrectly.

Although this example is a CAT designed for dichotomous classification, the same principles would be observed in an equiprecise CAT. The only difference would be in the termination criterion. Rather than ending the CAT when $\hat{\theta}$ was reliably below the cut score, an equiprecise CAT would end when the SEM associated with $\hat{\theta}$ fell below a prespecified value (e.g., .20). For example, had the test shown in Figure 2 been administered as an equiprecise CAT (rather than a dichotomous classification CAT), it would have required additional items to reduce the SEM associated with a computed $\hat{\theta}$ to a value of .20.

Is CAT Ready for Use in Counseling and Education?

Research support for CAT. Before a new technology such as CAT can be implemented, evidence must be available that supports its potential benefits and evaluates it effects on the psychometric characteristics of test scores that derive from it. Thirty years of research on CAT have provided ample evidence to support the theoretical benefits of CAT. This evidence has been derived from Monte Carlo simulation studies, post hoc or "real data" simulation studies in which conventional test data have been "re-administered" as CATs, and live-testing studies. Each of these types of studies has its advantages and limitations, and the resulting data from one type of study can be used to supplement and confirm some types of findings from other types of studies.

Early evidence of improved measurement precision (reliability) and validity (e.g., Johnson & Weiss, 1980; Kingsbury & Weiss, 1980) and large reductions in the number of items administered (typically 50% or more) without having an impact on the psychometric characteristics of test scores for CAT have been confirmed in a number of recent studies (e.g., Mardberg & Carlstedt, 1998; Moreno & Segall, 1997). Research that compares CATs and conventional tests also demonstrates substantial similarity between scores from the two procedures (e.g., Cudeck, 1985), although very high relationships should not be expected because conventional tests have limitations that are overcome by CAT. Recent research has concentrated on evaluating the potential benefits of technical improvements to CAT procedures, some practical issues (see below), and extending CAT to the measurement of personality variables (e.g., Reise & Henson, 2000; Waller & Reise, 1989) and attitudes (Koch, Dodd, & Fitzpatrick, 1990).

Some current CAT implementations. The feasibility of CAT has been supported in several large testing programs. One of the first was the national nursing licensure exam administered by the National Council of State Boards of Nursing (Zara, 1988). This testing program in paper-and-pencil format required examinees to spend 2 days taking a series of multiple-choice tests. Conversion of these tests to CATs (Zara, 1999) has reduced testing time to less than a day and provides immediate results for use in making licensing decisions.

A second major implementation of CAT took place in the achievement testing programs in the schools in Portland, Oregon. This CAT program focused on developing a longitudinal scale of achievement in several subject areas that spanned grade levels from early elementary through high school (Kingsbury, 1986). Once the item banks were calibrated using IRT methods, CAT procedures were implemented that allow students to be efficiently tested at periodic intervals, with each CAT efficiently and accurately locating each student's achievement level (Kingsbury & Houser, 1999).

Two other high-profile CAT programs include the Graduate Record Examination administered internationally by Educational Testing Service (e.g., Mills, 1999) and the Armed Services Vocational Aptitude Battery developed and delivered at testing centers throughout the United States by the U.S. Department of Defense (e.g., Moreno, 1997). In addition, five states in the United States have implemented statewide CATs for measuring student achievement (Olson, 2003, p. 12).

Some Current Operational Issues in CAT

Terminating a CAT. Some operational CAT programs are terminated by administering a fixed number of items or by imposing a time limit. Both of these termination procedures are used for the convenience of the test administrator. However, a test that is terminated for either of these reasons will not allow the CAT to continue until a CAT-based termination criterion can be implemented. If the CAT termination criterion is a specified minimum SEM, a prematurely terminated CAT will not result in equiprecise measurement, because the SEM does not decrease for all examinees at the same rate. Similarly, a CAT designed for equally confident classifications will, if terminated early, result in classifications of lower quality for some examinees. To obtain the maximum benefits of CAT, neither time limits nor a fixed test length should be imposed.

Content balancing. Much of the research and development in CAT has been done in the context of achievement testing. Although some achievement domains are both unidimensional and relatively homogeneous (i.e., they measure a single variable without substantial variation in content), some are relatively unidimensional but include two or more content domains. An example is arithmetic achievement at the elementary school level. The basic arithmetic operations (addition, subtraction, multiplication, and division) can be scaled on a single difficulty continuum, but they represent distinct operations for assessment purposes.

Because these operations can be scaled on a single difficulty scale, IRT procedures could be used to create an item bank for arithmetic achievement for use in a CAT. However, the difficulty differences among these operations would result in CATs that had different weightings of these operations across different examinees—students who had high ability would tend to get mostly division items and students who had low ability would receive mostly addition items. Thus, although all students would be measured on the same achievement scale, the content of their tests would differ across the four operations.

Several procedures have been proposed to achieve content balance among examinees in domains of this type (e.g., Kingsbury & Zara, 1991). These procedures modify the maximum information item selection procedure by also considering the content category of the items in the item selection process. Once an item is selected by maximum information at the examinee's current $\hat{\theta}$, its content classification is examined relative to target values specified in advance for each examinee. If the selected item represents a content area that is underrepresented at that stage in the examinee's test, the item is administered. If not, the item that provides the next highest information is examined relative to the content targets, and the process is repeated until an item from the appropriate content target is identified.

Clearly, by modifying the maximum information item selection procedure, content balancing reduces the efficiency of CATs. The result will be tests that are longer than they would otherwise need to be if content balance was not considered. In the context of measuring various traits of an individual to obtain quantitative data for use in counseling, content balancing is likely to be an issue only for CAT-based achievement tests. Measures of ability, personality, and preferences that have been constructed using classical item analysis methods (or IRT) will likely be relatively homogeneous in content and reasonably unidimensional and, therefore, will not require content balancing.

Multiple scales. Closely related to the issue of content balancing is the application of CAT to measuring instruments that use multiple scales to measure an examinee. Such instruments include ability test batteries (such as the Armed Services Vocational Aptitude Battery or the Differential Aptitude Tests), multiple-scale personality inventories, and attitude and preference scales constructed by classical test construction methods (e.g., the Multidimensional Personality Questionnaire; Patrick, Curtin, & Tellegen, 2002; Tellegen & Waller, 1992). For these types of instruments, the issue of content balance is achieved by treating each scale as a separate unidimensional variable and obtaining IRT parameters for CAT separately for each scale. Then, CAT can proceed separately for each scale to measure each examinee as well as possible on each scale. The result is, typically, a profile of scores for each examinee that can be used for counseling and other purposes.

When used for this type of measurement objective, CAT will provide highly precise and efficient measurements separately for each scale. However, the process of measuring an individual on multiple scales can be made even more efficient by an extension of the CAT procedure to the multiple-scale measurement problem.

Most test batteries or instruments with multiple scales result in scores that are intercorrelated to some degree. Ability test scores tend to correlate in the .30 to .50 range, and scales on personality and preference scales can have higher or lower intercorrelations, depending on the nature of the variables being measured by the scales. Because CAT can use differential starting values for beginning a test, the scale intercorrelations can provide information that can be used as starting values for tests after the first in a multiple-test application. Brown and Weiss (1977) proposed that scale intercorrelations among a set of achievement tests be computed using $\hat{\theta}$s from a test development group. The pair of scales with the highest correlation is chosen. Then, the multiple regressions of each scale as predicted from those scales are computed, and a new scale is added to the first two as the scale that can best be predicted from them. This process is then repeated adding one scale at a time. On the basis of these multiple correlations as each new scale is added, subtests can be ordered by how well they can be predicted from the other subtests. Finally, the multiple regression equations can be used to predict an examinee's initial $\hat{\theta}$ on each new test in the battery to be used as a starting value for that test.

This "inter-subtest" branching further enhances the efficiency of the separate CATs for each subtest by providing increasingly accurate starting values for each subsequent test in the battery. The result is further—and sometimes dramatic—reductions in the numbers of items needed to measure an examinee on multiple correlated traits. In the context of an achievement test battery, Brown and Weiss (1977; Gialluca & Weiss, 1979; Maurelli & Weiss, 1981) demonstrated that the length of later tests in a battery could be reduced by 80% or more of their full test length with no reduction in measurement precision.

Item exposure. As CATs are administered to groups of examinees, different items are taken by different examinees. Depending on the relationship between the trait distribution of the examinees and the information structure of the item bank, different items will be used (or "exposed") at differing rates. A number of procedures have been proposed to control for item exposure by not administering a selected item to an examinee if the probability of overexposure is high (e.g., Hetter & Sympson, 1997) or by modifying maximum information item selection to allow selection of items that are otherwise unlikely to be administered (e.g., Revuelta & Ponsoda, 1998). These procedures function similarly to content balancing procedures. That is, they modify the maximum information item selection procedure by constraining item selection to select items in order to control their probable exposure rate across a group of examinees.

Item exposure can be problematic in large testing programs or in some school settings when test scores are used to make decisions or judgments about examinees. In these cases, when there is an incentive for examinees to have access to items in the CAT item bank, the likelihood of an examinee's having prior access to the item (and the correct answer) is reduced by reducing the frequency of exposure of an item.

Similar to content balancing, item-exposure controls impose constraints on maximum information item selection and reduce the efficiency of CATs. Consequently, their use will result in longer tests than otherwise would be required. When tests are used for counseling and research purposes, however, there is rarely an incentive for examinees to improve their scores from prior knowledge of the correct or keyed answers to items that are likely to appear in a CAT. Hence, item-exposure controls are unlikely to be necessary in CATs used for these purposes, and unconstrained CATs can be used for maximally efficient and effective measurement.

CAT and the Worldwide Web. With the emergence of the Worldwide Web in the last decade, many tests of ability, personality, and preferences have been modified for delivery on the Web. Typically, a number of items are downloaded as a scrollable "page," the examinee answers the questions, and then returns the completed page through the Web. A long test might deliver several such pages.

Delivery of tests using the Web seems to be a logical next step from the earlier conversion of tests from paper and pencil to delivery by personal computer (PC) that occurred beginning in the 1980s. The PC testing movement, however, was based on a large body of research that supported the conversion of paper-and-pencil tests to PC administration by demonstrating that, for the most part, there were no effects on test standardization because of PC administration (e.g., Mead & Drasgow, 1993), except for tests that were primarily speed (as opposed to power) tests.

There has been almost no research to support the conversion of most tests to Web-based delivery, which can be quite different from PC-based delivery. In PC-based test delivery, the administration process is carefully standardized by software that will deliver a test in exactly the same way to each examinee. When tests are delivered by Web browsers, the variety of browsers and browser settings can potentially wreak havoc with test standardization, thus potentially invalidating test results. Without research to demonstrate the equivalence of Web-delivered tests, there is potentially great risk that the mode of administration might adversely affect the standardization of the instrument and affect the accuracy, validity, and utility of test scores.

Issues and recommendations concerning computer delivery of tests and Web delivery of tests have been addressed in a number of guidelines developed by several organizations, including the American Council on Education (1995); a joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999); the Association of Test Publishers (2000); and, most recently, the International Test Commission (2003).

The potential lack of standardization is even more likely to occur if a CAT is delivered over the Web. Because each item in an IRT-based CAT is selected on the basis of the examinee's scored answers to all previous items, computations must be implemented after each item response is received to select the next item, and the item bank must be available to deliver that item. It might, therefore, be tempting to deliver an item over the Web; send the answer back to the server for scoring, MLE, and selection of the next item based on item information; and then transmit the selected item to the examinee through the Web. This process would then need to be repeated for every item.

In addition to the delivery problems via the Web of conventional tests, this process would introduce an additional source of potentially negative influence on test scores—the response time of the Web. Although sometimes the Web responds quite quickly, there are other times when there are waits of several seconds or more. Such response times were typical in the 1960s and 1970s when electronic test delivery was attempted on time-shared computers. In most cases, the between-item delays were unacceptable and interfered with the standardization of the test-taking process, and time-shared delivery of standardized tests was generally abandoned until personal computers eliminated the time-shared delays. Item-by-item delivery of CATs via the Web would likely be a return to this approach of extremely unstandardized test delivery, thereby further compromising the utility and validity of test scores. McBride, Paddock, Wise, Strickland, and Waters (2001) provided a thorough discussion of a number of factors to be considered in the delivery of CATs based on standardized ability tests through the Web.

MEASURING INDIVIDUAL CHANGE WITH CAT

CAT is a viable technology that has potential to provide improved measurements, with substantially reduced testing times, that can be used in a variety of applications in counseling and education. In both of these areas, there is a need for measuring individual change. The counselor might be interested in whether counseling causes change in a student's behavior or adjustment. The educator might be interested in whether a student's level of achievement, understanding, or performance changes as a result of instruction or other educational interventions. Indeed, the No Child Left Behind Act (NCLB) of 2001 specifies that one of its goals is to "chart student progress over time." Yet another variation of CAT has been developed to accomplish this task.

Weiss and Kingsbury (1984) proposed an application of CAT that they called "adaptive self-referenced testing" (ASRT), which is designed to efficiently and effectively measure the progress of a single student over time. Like all CAT, ASRT begins with a set of test items that are calibrated on a variable of interest (e.g., an achievement domain such as reading or math) to reflect a wide range of the variable (e.g., from Grade 1 through Grade 12). A CAT is administered to a student and terminated with a fixed standard error. At a later date, another CAT is administered from the same set of items, but items previously administered to that student are not used. For the second (and subsequent) tests, however, the starting level for the CAT is based on the final $\hat{\theta}$ from the previous test, and the CAT is ended either when "significant" change has occurred or a fixed standard error has been reached. This process is repeated at later points in time, each time using the previous test's final $\hat{\theta}$ as the starting value for the next test. The result, as illustrated by Weiss and Kingsbury, is an individual profile of change (or lack thereof) that can be obtained with a minimum number of items for each student.

When measured change is identified by this procedure, the data can also provide information on when the change occurred for each student, thus identifying the points in the instructional or counseling process that had an impact on a student's measured levels on the trait. Data at each point in time can also be aggregated across students to track group progress. Recent research (VanLoy, 1996) has provided results that indicate that ASRT can better recover true change than can conventional procedures for measuring change.

Although NCLB has as one of its goals measuring progress over time, its current implementation seems to exclude procedures such as ASRT, or any procedures based on CAT. This is because the implementation of NCLB specifies that "states measure student performance against the expectations for a student's grade level" (Olson, 2003, p. 13), which is currently being strictly interpreted to mean that all students must be given the same test items, thus precluding the use of CAT in any form and especially a highly individualized procedure such as ASRT that involves repeated CATs. Only if this guideline is reinterpreted will NCLB be able to benefit from the many advantages that CAT can provide for the assessment of student progress at both the individual and group level. In the meantime, however, there are still a wide variety of testing applications in education and in counseling that can benefit from the improved and efficient measurement that will result from the application of CAT technology.

REFERENCES

American Council on Education. (1995). *Guidelines for computerized adaptive test development and use in education.* Washington, DC: American Council on Education.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Assessment Systems Corporation. (2001). The FastTEST Professional Testing System (Version 1.6) [Computer software]. St. Paul MN: Author.

Association of Test Publishers. (2000). *Guidelines for computer-based testing.* Washington, DC: Author.

Bayroff, A. G. (1964). *Feasibility of a programmed testing machine* (U.S. Army Personnel Research Office Research Study 6403). Washington, DC: U.S. Army Behavioral Science Research Laboratory.

Bayroff, A. G., Thomas, J. J., & Anderson, A. A. (1960). *Construction of an experimental sequential item test* (Research Memorandum 60-1). Washington, DC: Department of the Army, Personnel Research Branch.

Binet, A., & Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11,* 191–244.

Brayfield, A. H. (1961). Vocational counseling today. In E. G. Williamson (Ed.), *Vocational counseling, a reappraisal in honor of Donald G. Paterson*. Minneapolis: University of Minnesota Press.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Cleary, T. A., Linn, R. L., & Rock, D. A. (1968). Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement, 5,* 183–187.

Cleary, T. A., Linn, R. L., & Rock, D. A. (1969). An exploratory study of programmed tests. *Educational and Psychological Measurement, 28,* 345–360.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cudeck, R. (1985). A structural comparison of conventional and adaptive versions of the ASVAB. *Multivariate Behavioral Research, 20,* 305–322.

DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferguson, R. L. (1969). The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. *Dissertation Abstracts International, 30*(09), 3856A. (UMI No. 704530)

Gialluca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.

International Test Commission. (2003). International guidelines on computer-based and Internet-delivered testing (Draft Version 0.3). Louvain-la-Neuve, Belgium: Author.

Johnson, M. J., & Weiss, D. J. (1980). Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 16–34). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kingsbury, G. (1986). Computerized adaptive testing: A pilot project. In W. C. Ryan (Ed.), *Proceedings: NECC '86, National Educational Computing Conference* (pp. 172–176). Eugene: University of Oregon, International Council on Computers in Education.

Kingsbury, G. G., & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 93–115). Mahwah, NJ: Erlbaum.

Kingsbury, G. G., & Weiss, D. J. (1980). *An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests* (Research Report 80-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241–261.

Koch, W. R., Dodd, B. G., & Fitzpatrick, S. J. (1990). Computerized adaptive measurement of attitudes. *Measurement and Evaluation in Counseling and Development, 23,* 20–30.

Lofquist, L. H., & Dawis, R. V. (1969*). Adjustment to work: A psychological view of man's problems in a work-oriented society.* New York: Appleton-Century-Crofts.

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

Lord, F. M. (1971a). Tailored testing, an approximation of stochastic approximation. *Journal of the American Statistical Association, 66,* 707–711.

Lord, F. M. (1971b). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31,* 805–813.

Lord, F. M. (1971c). A theoretical study of two-stage testing. *Psychometrika, 36,* 227–242.

Mardberg, B., & Carlstedt, B. (1998). Swedish Enlistment Battery: Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB. *International Journal of Selection and Assessment, 6,* 107–114.

Maurelli, V. A., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

McBride, J. R., Paddock, A. F., Wise, L. L., Strickland, W. J., & Waters, B. K. (2001). *Testing via the Internet: A literature review and analysis of issues for Department of Defense Internet testing of the Armed Services Vocational Aptitude Battery (ASVAB) in high schools* (Report No. FR-01-12). Alexandria, VA: Human Resources Research Organization.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examination General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Mahwah NJ: Erlbaum.

Moreno, K. E. (1997). CAT-ASVAB operational test and evaluation. In W. A. Sands, B. K. Waters, & R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 199–205). Washington, DC: American Psychological Association.

Moreno, K. E., & Segall, O. D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169–179). Washington, DC: American Psychological Association.

No Child Left Behind Act of 2001, 107 U.S. C. § 1425 (2002).

Olson, L. (2003, May 8). Legal twists, digital turns: Computerized testing feels the impact of "No Child Left Behind." *Education Week's Technology Counts, 22*(35), pp. 11–14, 16.

Patrick, C. J., Curtin, J. J., & Tellegen, A. (2002). Development and validation of an abbreviated version of the Multidimensional Personality Questionnaire. *Psychological Assessment, 14,* 150–163.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347–364.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311–327.

Tellegen, A., & Waller, N. G. (1992). *Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire (MPQ).* Unpublished manuscript, University of Minnesota, Minneapolis.

Van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of item response theory.* New York: Springer-Verlag.

VanLoy, W. J. (1996). *A comparison of adaptive self-referenced testing and classical approaches to the measurement of individual change.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., et al. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption Scale. *Journal of Personality and Social Psychology, 57,* 1051–1058.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive?* (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

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

Zara, A. R. (1988). Introduction to item response theory and computerized adaptive testing as applied in licensure and certification testing. *National Clearinghouse of Examination Information Newsletter, 6,* 11–17.

Zara, A. R. (1999). Using computerized adaptive testing to evaluate nurse competence for licensure: Some history and a forward look. *Advances in Health Science Education Theory and Practice, 4,* 39–48.