Journal of Methods and Measurement in the Social Sciences
Vol. 2, No. 1, 1-23, 2011

Better Data From Better Measurements Using Computerized Adaptive Testing

David J. Weiss
University of Minnesota

The process of constructing a fixed-length conventional test frequently focuses on maximizing internal consistency reliability by selecting test items that are of average difficulty and high discrimination (a "peaked" test). The effect of constructing such a test, when viewed from the perspective of item response theory, is test scores that are precise for examinees whose trait levels are near the point at which the test is peaked; as examinee trait levels deviate from the mean, the precision of their scores decreases substantially. Results of a small simulation study demonstrate that when peaked tests are "off target" for an examinee, their scores are biased and have spuriously high standard deviations, reflecting substantial amounts of error. These errors can reduce the correlations of these kinds of scores with other variables and adversely affect the results of standard statistical tests. By contrast, scores from adaptive tests are essentially unbiased and have standard deviations that are much closer to true values. Basic concepts of adaptive testing are introduced and fully adaptive computerized tests (CATs) based on IRT are described. Several examples of response records from CATs are discussed to illustrate how CATs function. Some operational issues, including item exposure, content balancing, and enemy items are also briefly discussed. It is concluded that because CAT constructs a unique test for examinee, scores from CATs will be more precise and should provide better data for social science research and applications.

Keywords: Adaptive testing, computerized adaptive testing, conventional tests, item response theory, measurement error, off-target tests

A considerable amount of social science data is obtained using methods of psychological measurement. These methods include tests, inventories, and scales used to measure ability, achievement, proficiency, personality, attitudes, and a variety of other variables of interest to researchers and practitioners in psychology, education, sociology, political science, and other disciplines and applications. The majority of these instruments were developed by classical test theory methods.

Classical test theory (CTT; e.g., Gulliksen, 1950; Allen & Yen, 1979/2002) is designed for the development of conventional tests—measuring instruments that use a fixed set of questions/items that are selected based on data from a target group of respondents. A trial set of test items is administered to the group and the resulting data are used for an "item analysis," in which two types of statistics are typically computed for each item: (1) item difficulty, defined as the proportion of respondents who answered the item in the keyed (or correct) direction, or for a rating scale type of item the mean total score for a given item response; and (2) item discrimination, defined as the correlation of the item response with total score on the scale to which the item belongs. The next phase of item analysis typically is to select items that have item difficulties (or means) near the center of the range of item difficulties. For dichotomously scored items (correct/incorrect, keyed/non-keyed), this means selecting items with proportion correct near .50; for rating scale items, it means selecting items with mean scores near the center of the rating scale weight range. Items with extreme means or proportions are usually deleted from the measuring instrument. The next step in an item analysis is to delete items that have low correlations with total scores.

The objective of these two steps in a item analysis for conventional tests is to increase the internal consistency reliability of the scale or instrument, as reflected in indices such as Cronbach's alpha (Cronbach, 1951). This type of reliability is increased by eliminating items with extreme difficulties, because these items have low variance and by eliminating them the variance of the total score is increased, since the total score is based on all items; increasing the variance of total scores relative to the number of items increases reliability. Reliability is also increased by eliminating items with low correlations with total score, because internal reliability coefficients are proportional to the average item intercorrelation and the item-total correlation is proportional to the average correlation of an item with the other items. The process of refining such a measuring instrument involves recomputing reliability as these two steps are implemented and ending the instrument refinement process when either a sufficiently high level of reliability is reached, or eliminating additional items results in only trivial increases in reliability. Reliability in CTT can be thought of as "precision" of measurement since a complementary function of reliability can be expressed as "standard error of measurement." Reliability in CTT is computed for a specified set of test items from data collected on a particular group of examinees. It is a single value (as is the standard error of measurement derived from it) for that set of items measuring that group of individuals.

A Perspective From Item Response Theory

Although these instrument development procedures have been in use for almost 100 years, their full implications with respect to the nature of the resulting measurements were not evident until the more modern methods of item response theory (IRT) became available in the mid 1970s. IRT is a family of mathematical models that formalize how individuals respond to items in psychological measuring instruments (de Ayala, 2009; Embretson & Reise, 2000). These models include models for dichotomously scored items as well as rating scale items and other types of items that result in multi-category (polytomous) responses. IRT includes some concepts that are not part of classical test methods, and some of these concepts can be applied to describe the effects of constructing instruments using CTT test construction procedures.

One of these concepts is test information. Information in IRT replaces the concept of reliability used in CTT. It can also be interpreted as "precision" of measurement, but it differs in several ways from CTT's "precision"—higher information means more precision in differentiating two closely contiguous levels of the variable being measured. (In IRT, the variable is generally referred to as a "trait" in a very broad sense—it represents any unidimensional variable, whether ability, aptitude, attitude, or personality variable, and is typically symbolized with the Greek letter θ). Although CTT reliability is a constant for a set of test items applied to a group of individuals—every score computed from that set of items has the same precision or error of measurement—information in IRT is a function that allows precision to vary at different levels of θ. Similar to reliability in CTT, test information in IRT can be converted to an error of measurement, but that error of measurement is a function of θ level, not a single value. The standard error of measurement (SEM) function is obtained by taking the reciprocal of the square root of information at each value of θ. Thus, in IRT there is not one SEM

for a given set of items but rather an infinite number, potentially a different value for each potential level of θ based on a given set of items.

Figure 1a shows the test information function for a 50-item typically constructed conventional test. The 50 items are highly discriminating items that all have item difficulties around .50, resulting in a "peaked" test characteristic of conventional test development procedures. As Figure 1a shows, the information is high and maximum at the center of the θ scale (θ = 0.0) and drops rapidly as θ moves away from the center. Figure 1b shows the conditional SEM function for the same test. As the figure shows, the SEM is smallest (about 0.12) at the center of the θ distribution and increases rapidly for examinees with θs above or below the mean, becoming greater than 0.50 at $\theta = \pm 1.8$. These observations show that conventionally constructed measuring instruments are designed to measure well at a point (typically the mean of the score distribution) but, because they are based on a fixed set of items selected to measure around that point, they measure increasing poorly for individuals whose scores deviate from that point, with levels of measurement error increasing rapidly with increasing distance from the score mean. Thus, scores near the mean of the distribution are relatively precise, but scores away from the mean have considerable error associated with them.

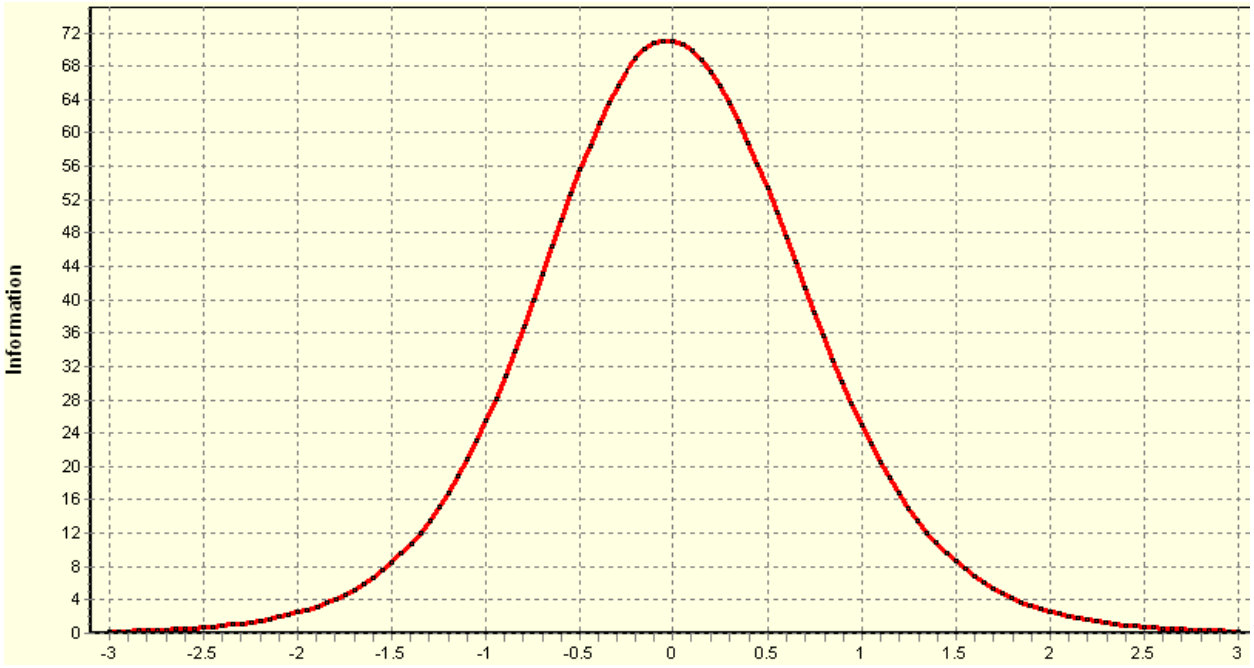
Effects of Measurement Error on Score Variability

Although not widely understood, measurement error, even in CTT, operates to artificially increase score variability, but the increase is due to random factors. Since random variability (i.e., measurement error) by definition is "noise," the increased variability in CTT test scores can serve to lower the correlation of test scores with other variables and also affect the results of other statistical analyses using error-laden scores. Random data will not correlate with other variables (as recognized by CTT's "correction for attenuation"), thus reducing the predictive validity of CTT scores. Similarly, random variability increases the "error" terms in tests of mean differences and related analyses, also reducing the ability of the scores to reflect mean differences in research studies.

Figure 2 shows the effect of error of measurement from conventional tests on test scores as the examinee's true trait level deviates from the point at which a conventional test is peaked. These data were derived from a small monte-carlo simulation study using a 50-item peaked conventional test, similar to that shown in Figure 1, administered to examinees at trait (θ) levels distributed closely around the point on a standard score scale (mean = 0.0, standard deviation = 1.0) where the test was peaked (θ = 0.0), and for examinees whose true θ levels deviated from the test, at θ = .60, 1.2, 1.8, and 2.4. These results are contrasted with those from a computerized adaptive test (CAT; discussed below) that also administered a 50-item test selected dynamically for each examinee from a larger item bank. Number-correct scores on the conventional test were converted to the IRT θ (standard score) metric so that they could be compared with the true θs and the θ estimates from the CATs.

Figure 2a shows the effects of error of measurement on mean test scores expressed as bias—the mean difference between estimated scores and true scores. When the examinee θs are clustered around the value where the test is peaked (θ = 0.0), scores from the peaked conventional tests (red bar) are unbiased. This is also the point at

a. Test Information Function



b. Test Standard Error of Measurement Function

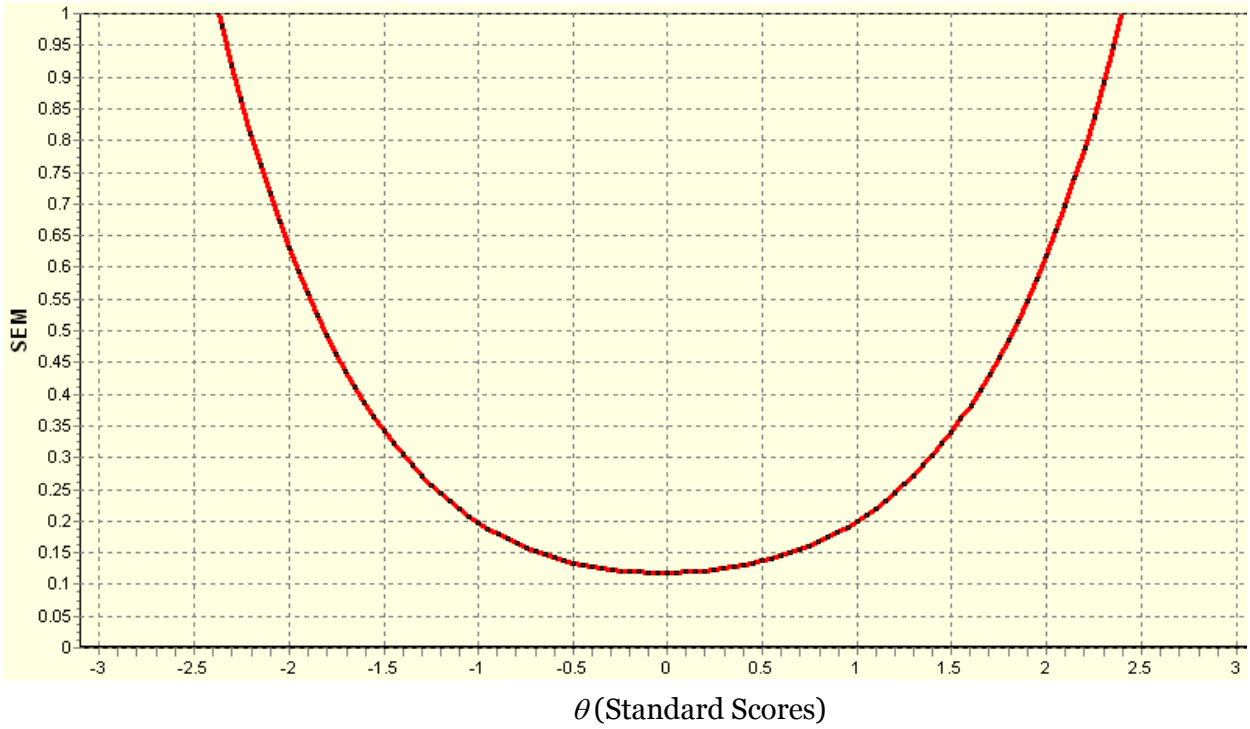


Figure 1. IRT Functions for a Peaked Conventional Test

which test information is maximum (Figure 1a) and test standard error (Figure 1b) is minimum. When mean θ = 0.6, peaked test scores are nearly unbiased. However, as test information falls off around θ = 1.20 and the test SEM doubles, mean bias of the peaked test scores is 0.20. For mean θ = 1.80 (almost two SDs above the mean), the SEM has quadrupled and the true mean is overestimated by 0.50 θ units. By contrast, the green bar show that the CAT θ estimates were essentially unbiased regardless of the θ levels of the examinees.

Figure 2b illustrates the effects of conditional errors of measurement on the SDs of the converted number-correct scores. The first bar in each set in the figure is the SD of true θ (a constant value of about 0.13, approximately equivalent to the test SEM at θ = 0.0); the second bar is the SD of θ estimates for CAT; and the third the SD of converted number-correct scores from the conventional test. When mean examinee θ matched the difficulty of the test (θ = 0.0), both the conventional test and the CAT had essentially equal SDs that slightly over-estimated true θ. As θ deviated from 0.0, the SDs for the CAT remained essentially equal, reflecting the constant error of measurement characteristic of CATs. By contrast, the SDs of the peaked conventional test increased with increases in θ. At θ = 1.20, the SD of number-correct scores was 0.62—almost five times the true SD. At θ = 1.80, the SD of observed number-correct scores was 0.76—almost six times the true SDs. The decline in SD at θ = 2.40 is due to a ceiling effect on the number-correct scores.

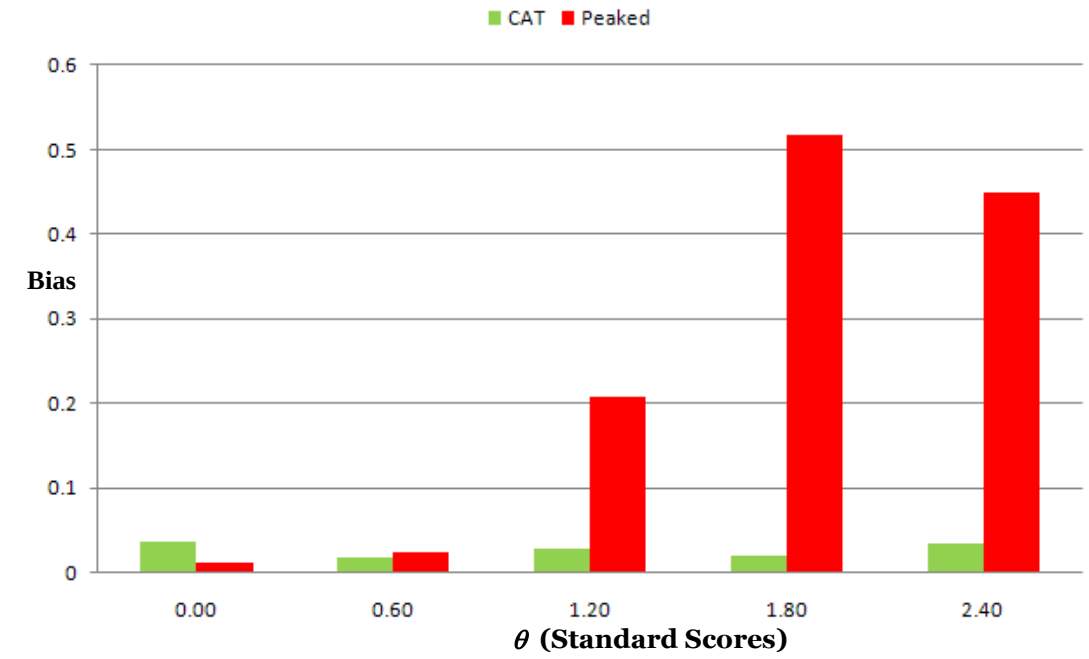
These results show substantial bias in number-correct test scores and significant artifactual increases in score variability from conventional tests when administered to examinees whose trait levels do not match the difficulty of the test. These spurious effects increase as examinees deviate from the point at which the test is peaked. In a given sample of examinees, however, the actual effects of errors of measurement of this type will be unknown because (1) number-correct scores are not error-free indicators of true trait levels, and (2) the true trait distribution is unknown. Embretson (1996) and Kang and Waller (2005) also demonstrated, in computer simulation, the negative effects on conventional test scores of "test inappropriateness"—the "off-target" use of conventional tests—in the context of detecting interactions in ANOVA analyses and in moderated multiple regression.

It is obvious that there can be substantial measurement error in test scores from peaked conventional tests and that the error can have serious detrimental effects on conclusions drawn from the use of those measurements. By contrast, Figure 2 shows that CATs are not susceptible to these effects.

A Real-Data Example

Only one study appears to have examined, using real data, the effects of the more precise scores of CAT in comparison to those of conventionally administered tests. Gibbons, Weiss, et al. (2008), developed a CAT version of a psychiatric scale—the Mood and Anxiety Spectrum Scales (MASS)—designed to measure mood and three other important psychiatric variables. The MASS, developed using CTT procedures applicable to developing personality inventories, consists of 626 yes/no items that result in an overall score and four subscores. The authors applied CAT to the MASS using a bifactor CAT algorithm with maximum information item selection, Bayesian θ estimation, and a

a. Mean Bias



b. Standard Deviations (SD)

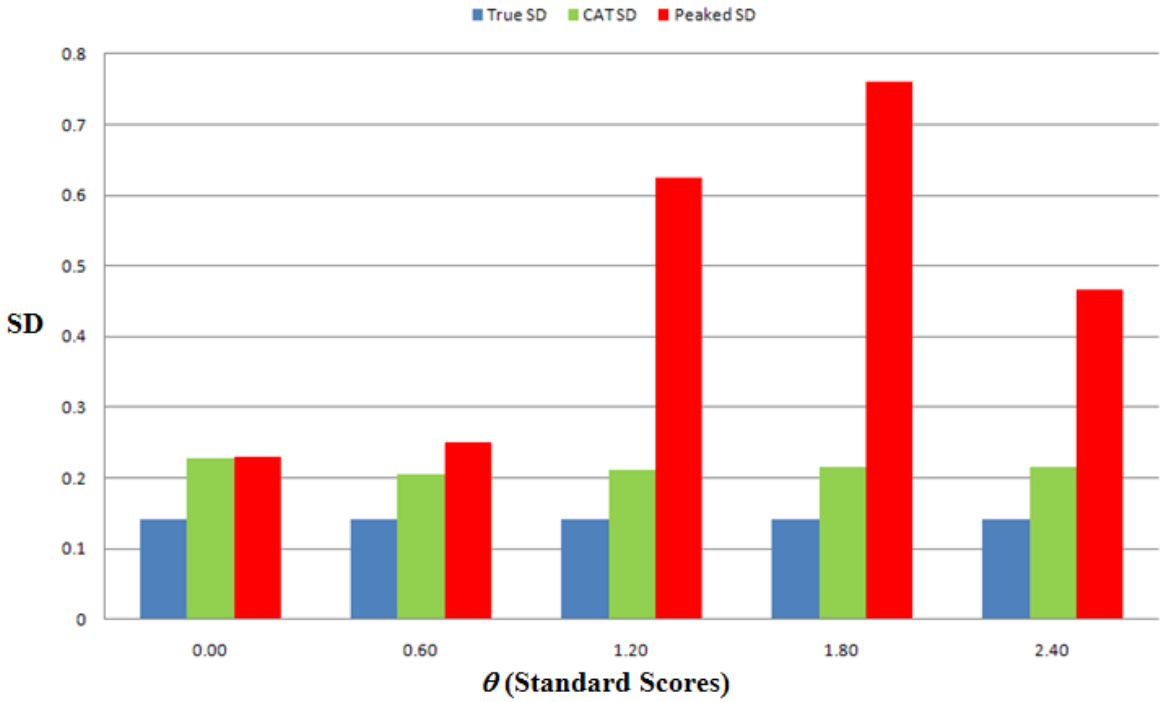


Figure 2. Converted Number-Correct Scores for a Peaked Conventional Test and CAT θ Estimates at Five Levels of θ

SE = .30 termination criterion for the general factor. Results indicated a 95% reduction in scale length for the general scale as well reductions of 85% or more for each of the four subscales.

Scores on the Mood scale were contrasted for two diagnostic groups—with and without independently determined bipolar disorder. Conventional scoring of the 161-item Mood scale resulted in a significant difference between the mean scores of the two groups: $t = 3.20$, $df = 154$, $p < .003$, an effect size of .63 SD units. By contrast, the CAT required an average of only 27 items (an 83% reduction in scale length) and resulted in $t = 6.00$, $df = 154$, $p < .001$ for an effect size of 1.19 SD units. Thus, the CAT scores identified an effect almost twice as large as that of the conventional scores, as a result of the greater precision of θ estimates—due to a reduction in error variability—obtained by the CATs.

Basics of Adaptive Testing

Contrary to popular belief, adaptive testing is not new—although CAT is obviously relatively recent. The basic principles of CAT were articulated and implemented by Alfred Binet in 1905 in what later became the Stanford-Binet IQ test (Binet & Simon, 1905). By contrast, the conventional fixed-form paper-and-pencil test was not widely implemented until around 1918 when it was used to efficiently screen military results for the U. S. armed forces in World War I (Dubois, 1970). Its use then expanded dramatically over the years until the paper-and-pencil test dominated psychological and educational testing for most of the twentieth century.

In intelligence testing, the Stanford-Binet adaptive test has been considered the "gold standard" against which the vast majority of subsequent intelligence tests have been judged. Binet's test, individually administered by a trained psychologist, incorporates all the characteristics of current adaptive tests, but obviously in a different form than contemporary CATs. An adaptive test is comprised of five characteristics that differentiate it from conventional tests:

1. It is based on an item bank with test items of known psychometric/statistical characteristics. The item bank is typically a wide-ranging bank that covers a defined range of the trait to be measured.
2. Test administration can use information available on a given examinee to select a starting point for the examinee in the item bank—not all examinees are required to start with the same item or item set.
3. Items are scored as they are administered and a test score can be derived from different subsets of items given to different examinees.
4. Some type of rule is used to select subsequent items based on an examinee's scored responses to previous items.
5. An examinee's test is ended when a prespecified termination criterion is reached—a fixed number of items is not necessarily administered to every examinee.

As a result of the last four characteristics, an adaptive test is an *individualized* test. Examinees need not start with the same items, each examinee can receive different subsets of items, and examinees can receive different numbers of items from the bank.

An adaptive test is dynamic—it adjusts the difficulty of the test administered to the trait level of the examinee as the test is being administered.

Binet's Adaptive Test

Figure 3 is a schematic representation of Binet's adaptive test administered to a single examinee. The item bank for this hypothetical test consists of 210 items organized in 21 "mental age" levels (with 10 items per level) at half-year intervals from 5 to 15. Binet defined the 'mental age" of a test item as the chronological age of a group of examinees who answered his free-response test questions correctly 50% of the time. Thus, for example, if approximately 50% of a group of 10-year-old children answered a given test item correctly, that item would be placed in the 10-year old "mental age" group of items; the same item might be too difficult for 9-year-old children (only 35% might correctly answer it) or too easy for 11-year-old children (85% might correctly answer it).

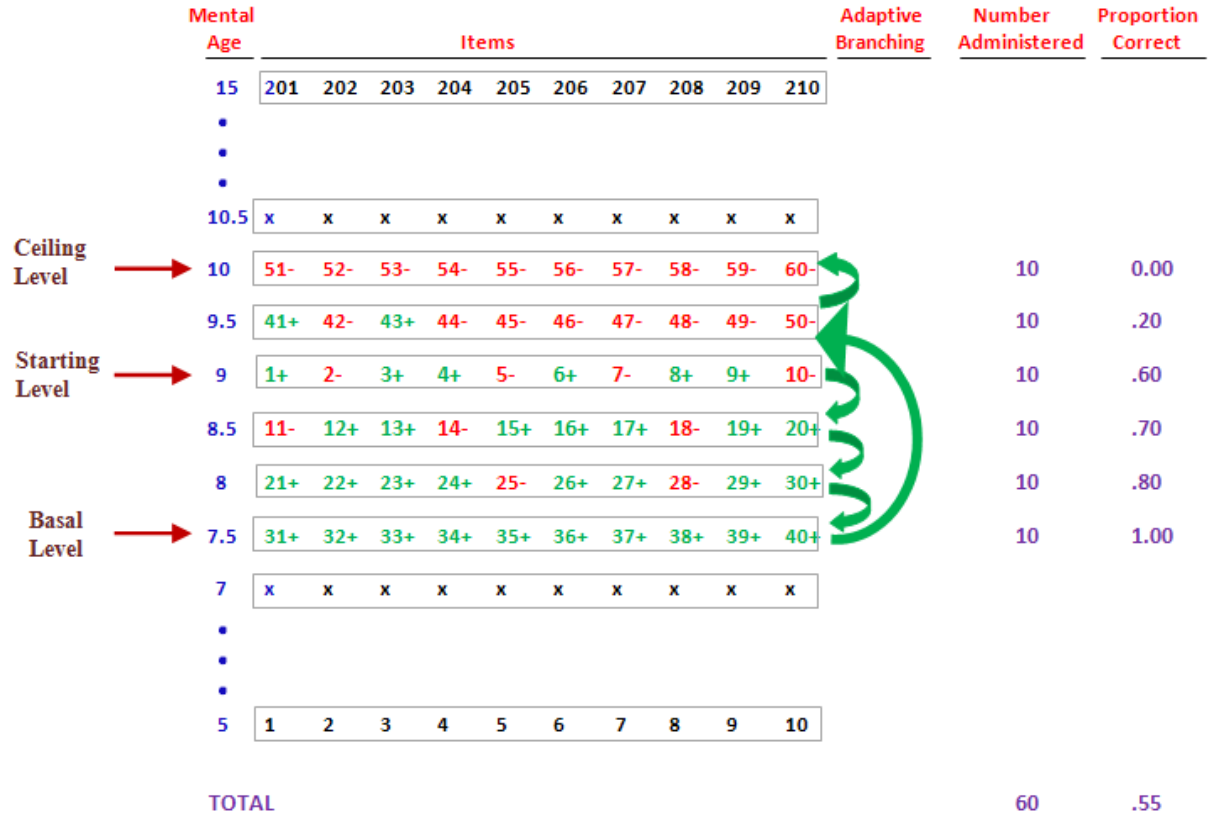


Figure 3. A Schematic Representation of a Binet Adaptive Test

Given an item bank structured in this way, the first step in adminstering the test to a given child is to select a starting level to begin the test. Similar to today's much more sophisiticated adaptive tests, a Binet test allows the use of prior information to select the first item for the test—this is the first aspect of adapting the test to a given examinee. If the examiner knows something about the child that is relevant to her/his probable

performance on the test, that information can be used to select a starting level for the test. In this hypothetical example, although the child's chronological age was 8 years, the examiner might have information to indicate that the child is thought of as "bright," so the test was begun at the 9-year mental age level; in the absence of that information, the test would likely have started at the 8-year level.

Once the starting level is selected, the first item at that level is administered and immediately scored, in this case correct (1+). Succeeding items at the current level are similarly administered and scored, with a result at Mental Age 9 of six out of 10 items correctly answered for a proportion correct of 0.6. At this point, the examiner is faced with a second adaptive decision: Should the test continue by administering easier items in a search for the child's "basal level"—the mental age at which the child answers all the items correctly—or should more difficult items be administered to attempt to determine the "ceiling level"—the mental age level at which all items are answered correctly?

In this example, the examiner chose to identify a basal level first, so items at the next lower difficulty level (Mental Age 8.5) were administered. Each item was scored immediately by the examiner, and the result was a proportion correct of 0.70. Because this result did not identify a basal level, the next level of easier items (Mental Age 8.0) was selected and those items administered and scored with a resulting 0.80 correct. Finally, after further adapting the level of difficulty to the child being tested by dropping down one more level of difficulty, the child correctly answered all ten items at Mental Age 7.5 and a basal level was established. This result indicated to the examiner that it was not necessary to administer any easier items, so all items at Mental Age 7 and below (a total of 50 items) were skipped for this child.

Having identified the child's basal level, the examiner then proceeded to identify the ceiling level—the level of difficulty that identifies the child's upper limit of ability. Since all items at Mental Ages 7.5 through 9 had been administered, the test was adapted by administering items at the next available level above Mental Age 9. Thus, the ten items at Mental Age 9.5 were administered and scored, with a resulting proportion correct of 0.20. Because this level of performance did not identify a ceiling level, items of Mental Age 10 were administered. The resulting 0.0 proportion correct identified the ceiling level, and the remaining 100 more difficult items in the bank were not administered. Finally, a "mental age" score is computed by a weighted average of the mental ages of correctly answered items, this result is divided by the child's chronological age, and then is multiplied by 100 to arrive at the child's "I.Q."

The adaptive procedure incorporated into a Binet-type test essentially identifies the effective range of item difficulty for each examinee. Examinees who are capable of answering more difficult items will be administered those items; examinees who are unable to answer difficult items will be given easier items. Thus, different examinees will receive different subsets of items drawn from the pre-calibrated item bank and, with the exception of incorrect starting levels, will receive a minimum number of items that are too difficult for them or those that are too easy. As a consequence, adaptive testing is efficient—it administers only those items necessary to measure a given examinee and eliminates most items that provide little or no information about the examinee's ability level. The efficiency is illustrated in this example: Without using an adaptive procedure, all 210 items would have had to be administered to this child to obtain an adequate measure of ability. But the adaptive testing procedure accomplished the measurement objective in only 60 items, eliminating 50 items that were too easy for the

child (and therefore provided no information about her/his ability level) and 100 items that were too difficult and similarly uninformative. The resulting 60-item test achieved a 71% reduction in test length from administering the entire item bank as a fixed-length conventional test. One important characteristic of an adaptive test is that the use of variable termination criteria will result in different length tests for different examinees. Figure 4 shows, for three different students whose mental age scores were similar, a schematic of the number of items administered to each student. Student A, whose hypothetical response record is shown in Figure 3, received items that ranged from Mental Age 7.5 through 10; Student E received items only for Mental Ages 8.5, 9, and 9.5; and Student F received items from Mental Ages from 6.5 through 11.5. Clearly, Student E is measured with the most precision (his/her score is more certain) than either of the other two students, and Student F's mental age score will be the least precise—this student is interacting with the item bank in a manner different from the other two students. Thus, an adaptive test can yield not only a score estimate, but an indicator of the precision associated with that score.

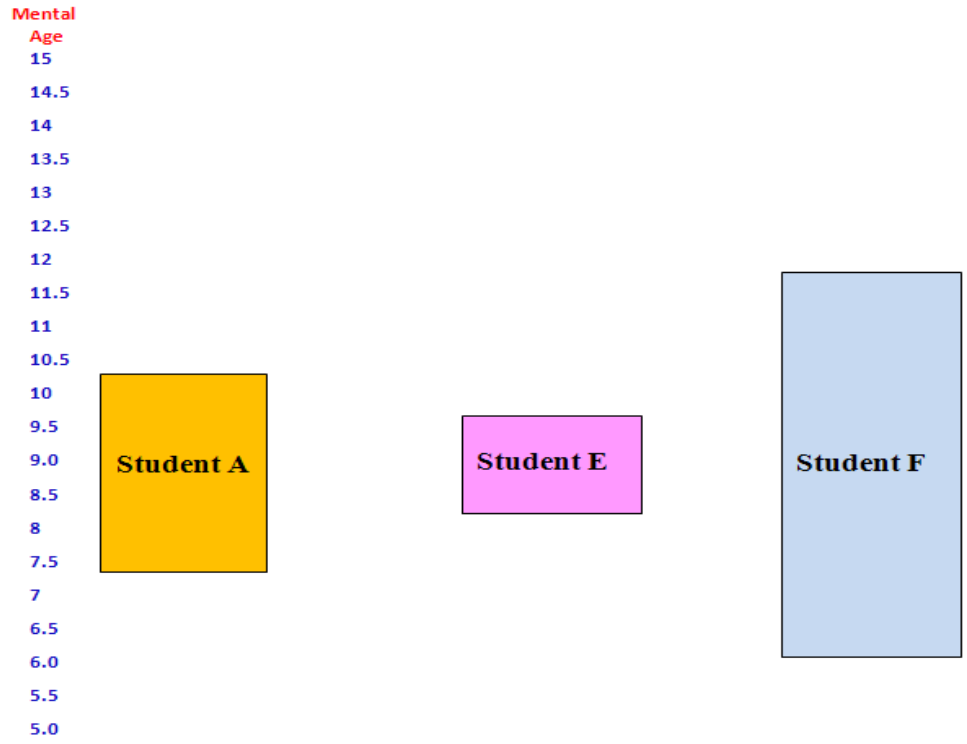


Figure 4. Ranges of Items Administered to Three Students on a Binet-Type Test

Although Binet's test has been extremely useful for the purposes for which it was developed, it is not without its problems. First, the test is administered individually by a psychologist, making its widespread application limited due to cost and the relative unavailability of trained administrators. Second, although it makes efficient use of an item bank, it is still inefficient in some respects. For example, it requires that all items at a given mental age be administered before an adaptation occurs, making it only partially adaptive. As a consequence, if a test administrator underestimates (or

overestimates) an examinee's actual mental age by several levels, it might require (in the example shown) several blocks of ten items before a basal or ceiling level is obtained, thus unnecessarily lengthening the test and reducing its efficiency. Third, although the test yields a subjective indicator of the precision of a test score for a given examinee, it has no explicit mechanism for controlling score precision.

Fully Adaptive Computerized Adaptive Testing

The problems with Binet's tests were resolved with the introduction of computers into the testing process in the early 1970s (e.g., Weiss, 1973) and have been refined into highly efficient and effective procedures for measuring individuals. Modern fully adaptive CAT is based on item response theory (IRT), a family of mathematical models that describe how examinees respond to test items of various kinds (e.g., DeAyala, 2009, Embretson & Reise, 2000). These models can be applied to items that are scored correct or incorrect (or "keyed/not-keyed"), items scored by assigning partial credit to responses to multiple-choice items, or to rating scale items used to measure a wide variety of attitudes, perceptions, and personality variables. By combining IRT with the test delivery capabilities of computers, fully adaptive CAT allows item responses to be scored immediately and adaptation to occur after each item is administered. IRT also allows scores, and associated error bands for those scores, to be calculated after each item is administered and that information can be used to select the next item or to end the test for a given examinee.

Item bank. As with the Binet test, the first step in implementing an IRT-based CAT is to develop an item bank with psychometric data on the items. In contrast to the Binet bank, however, a fully adaptive CAT item bank is not structured, although some forms of partially adaptive CATs use structured item banks based on IRT item data (e.g, Chang, Qian, & Ying, 1999; Chang & van der Linden, 2003; Zenisky, Hambleton, & Luecht, 2010). In IRT, test item "difficulty" and "discrimination" are defined differently than they are in CTT, but for purposes of CAT are combined into an *item information function* (IIF). Similar to test information functions, like that shown in Figure 1a, item information is a function that reflects how precisely a single test item measures at various points along the θ continuum. Higher information indicates greater precision and low information indicates a lack of precision. Figure 5 shows IIFs for four items. The location of the curve along the θ axis reflects the difficulty of the item. Thus, Item 1 is the least difficult because it is located at the lower (negative) end of the θ continuum, and Item 4 is the most difficult. The height of the IIF at its maximum reflects the discrimination of the item—how well it differentiates between examinees whose true θ levels are close together; Item 1 is the most discriminating and Item 4 is the least. A CAT item bank for measuring a particular variable might have as many as 200 or more items, and IIFs are calculated for each item.

Starting a CAT. The second step in implementing a CAT is to identify some rule for starting the test for an examinee. As with the Binet test, the first item to be given to an examinee can be based on prior information about the examinee, it can be the same for all examinees, or it can be randomly selected from a set of items within a limited range of the trait continuum. Prior information, if accurate, will increase the efficiency of a CAT; on the other hand, in a fully adaptive CAT, incorrect prior information will

reduce its efficiently only marginally since (as will be shown below) CATs can recover quickly from incorrect starting points. Consequently, randomly selected starting items will have little effect on CATs and will serve to reduce "item exposure," which can be important in CATs that are used to make high-stakes decision about examinees. If a constant starting item is used, all examinees will receive the same first item and will also see a restricted range of items for the first few items in the test.

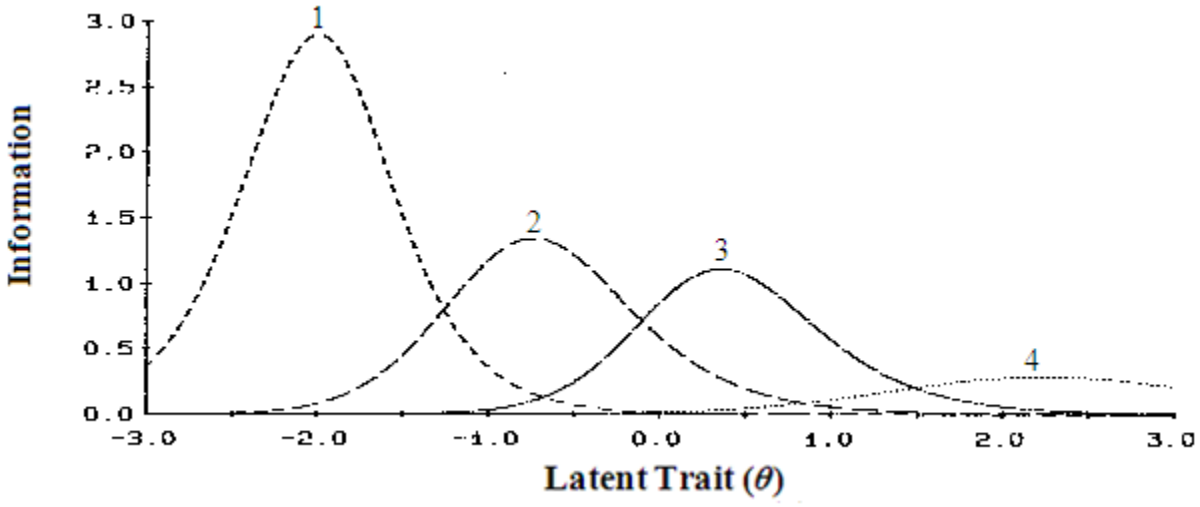


Figure 5. IFFs for Four Dichotomously Scored Items

Estimating θ. In fully adaptive CAT, an examinee's θ level is estimated after each item is administered and immediately scored. Using IRT θ estimation methods, an examinee's θ level can be estimated after a single item is administered or after two or more items are administered. The general method for estimating θ in IRT uses maximum likelihood estimation (de Ayala, 2009, pp. 347-355). However, when only one scored item response is available at the beginning of a CAT (or, if several items have been administered and they have all been answered either correctly or incorrectly) the maximum likelihood procedure must be modified temporarily in order to obtain a finite θ estimate. This modification, which temporarily assumes that θ for a group of examinees is normally distributed, is called Bayesian estimation (de Ayala, pp. 77-79). The Bayesian θ estimate after the first item is administered is then used to select the next item for that examinee (although sometimes an arbitrary increase or decrease in θ is used in place of a Bayesian estimate). If the examinee correctly answers the first item (or answers in the keyed direction if there is no "correct" answer), the examinee's θ estimate will increase; if the answer is not correct (or keyed), the θ estimate will decrease.

The θ estimation process continues as new items are selected and scored, with θ estimated anew after each item response. Once a mixed response pattern is obtained (e.g., 01, where 1 is a correct/keyed response and 0 is an incorrect/not-keyed response) the normal distribution assumption is no longer required and non-Bayesian maximum likelihood estimation is used. One major advantage of maximum likelihood estimation of θ is that it takes into account all the information in an examinee's responses in conjunction with all the information available on each test item. Thus, for example, if

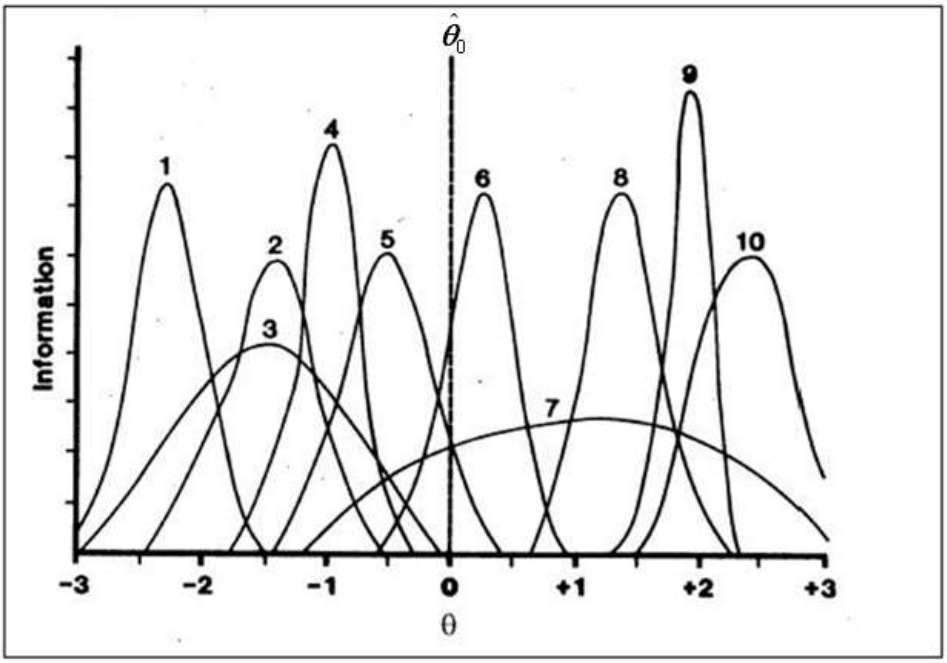
an examinee correctly answers a difficult item, his/her θ estimate will increase more than if he/she correctly answers an easier items. Similarly, if an examinee incorrectly answers an easy item, his/her θ estimate will decrease more than it he/she had incorrectly answered a more difficult item. As a consequence, IRT scoring will provide different θ estimates for four items answered 1100 versus the same items answered 0011—the number-correct score cannot differentiate these two examinees, but IRT θ estimation will. A second advantage of IRT θ estimation is that it will provide a standard error of the θ estimate each time θ is estimated. These empirical standard errors reflect the confidence that the test user can have in a given θ estimate and can be used to end a CAT for an examinee.

Item selection. As indicated, fully adaptive CAT is differentiated from other forms of CAT in that items are selected, administered, and scored one at a time, θ is estimated after each item is given, and a new item is selected to continue the test. Item selection is based on the IIFs for all the unadministered items in the bank. At each stage of the CAT—i.e., after each θ estimate—the next item to be selected is the unadministered item at the examinee's current θ level that has the highest level of item information; this process is known as *maximum information item selection*. Thus, for a given θ estimate, in effect the information available from each item given that θ estimate is computed and the previously unused item (for that examinee) with the highest information is selected and administered. As it turns out, that item is the item that will maximally reduce the error of the next θ estimate obtained after that item response is scored. This property of the CAT process typically results in two outcomes: (1) differences in successive θ estimates tend to decrease as more items are administered, and (2) the standard errors associated with the successive θ estimates will tend to decrease and converge throughout the CAT.

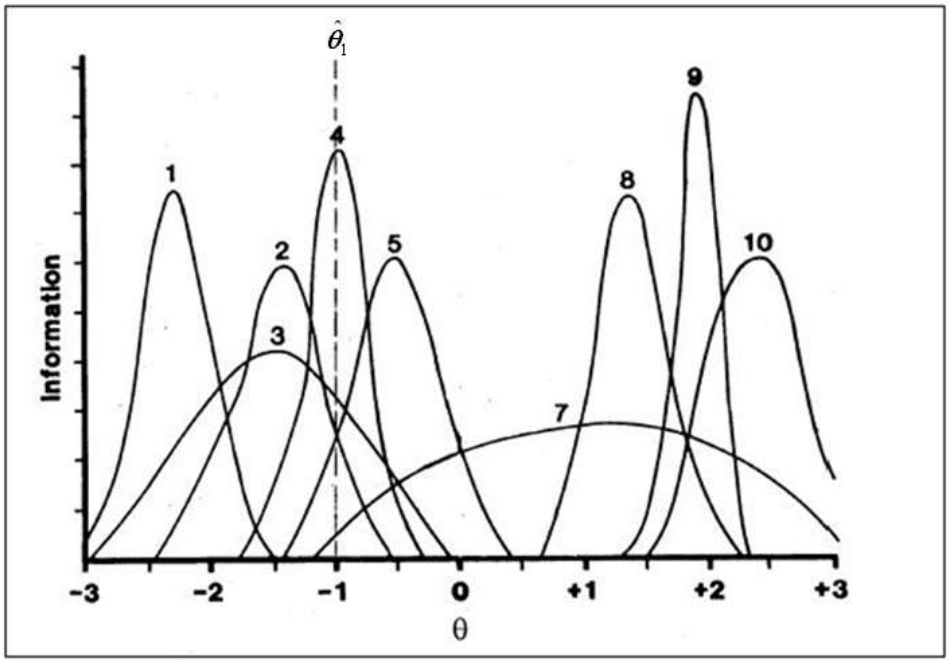
Figure 6 illustrates maximum information item selection for a hypothetical set of 10 items (obviously a real item bank will have many more items). Item 9 is the most discriminating item (its IIF has the highest peak) and Item 7 is the least discriminating item (its IIF is the flattest); Item 10 is the most difficult (it provides information for high θ examinees) and item 1 is the least difficult (it differentiates only among low θ examinees). The vertical dashed line in Figure 6a shows the starting θ estimate ($\hat{\theta}_0$) of 0.0 at the beginning of the CAT. Of the three items that provide non-zero values of information at θ = 0.0 (Items 5, 6, and 7), Item 6 has the maximum amount of information, so that item is selected from the bank, administered and scored, and θ is estimated (in this case using the Bayesian prior distribution), resulting in $\hat{\theta}_1$. In this example, $\hat{\theta}_1$ = –1.0, which resulted from an incorrect response to Item 6. Figure 6b shows the item bank after Item 6 has been removed and the vertical dashed line indicates that five items—Items 2, 3, 4, 5, and 7—had non-zero information at $\hat{\theta}_1$, and that Item 4 had the highest IIF at that point. Therefore, Item 4 is displayed, the answer recorded and scored, and θ is re-estimated. The figure shows an increase in $\hat{\theta}$ (resulting from a correct response) to about $\hat{\theta}_2$ = –0.50, where Figure 6c shows that Item 5 provides maximum information. The process continues—the selected item is administered, scored, θ is re-estimated (using all the item responses available), and the next item

providing maximum information at the current θ estimate is administered and removed from the bank—until a termination criterion is reached.

a. 10-Item Bank at the Start of a CAT



b. One item Administered



c. Two Items Administered

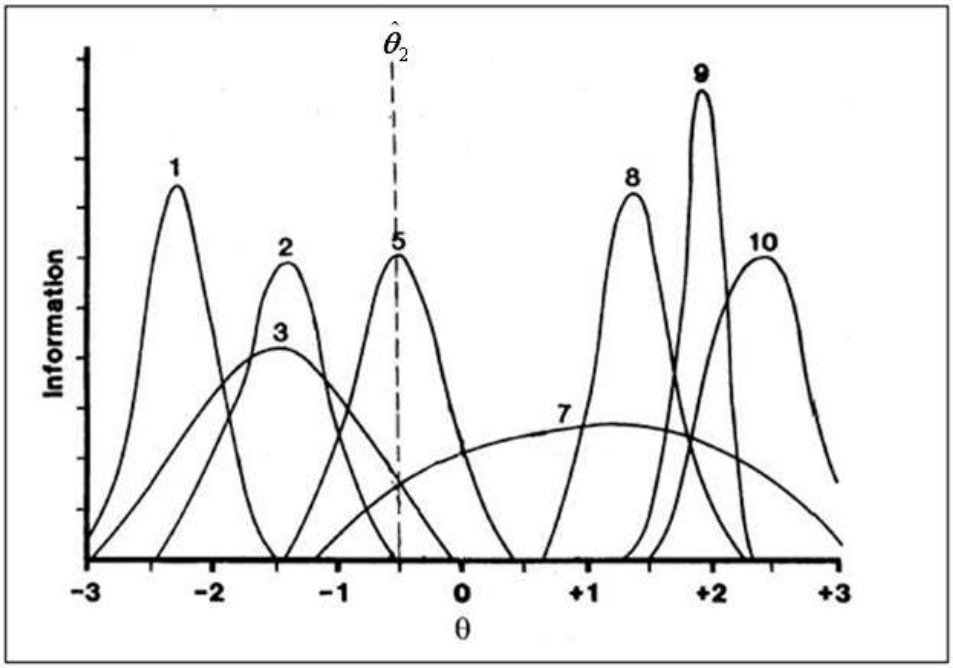


Figure 6. Maximum Information CAT Item Selection

Ending a CAT. A properly implemented CAT uses a variable termination criterion consistent with the purpose of testing. CATs can be used for a number of purposes, including:

1. Measuring individuals to obtain scores that are used to evaluate the examinee's level of functioning on some trait of interest. Such scores might be used for counseling purposes, clinical evaluation, in schools, in a variety of other settings where an individual's level of functioning is important information, or in research studies. So that scores of individuals will be comparable in terms of quality, or to minimize the kinds of error variability described above, such scores should be of equal precision across individuals.
2. Classifying individuals based on one or more score cutoff values. In this case, test scores are used to determine if an examinee has mastered or not mastered a body of knowledge, has passed or not passed a course of study, or qualifies or does not qualify for a particular employment or educational opportunity based on knowledge or skills. Classifications can also be made based on multiple score cutoff values, such as in the assignment of school grades or proficiency categories.
3. Measuring individual change, growth, or decline (or lack thereof). In this type of application, precise scores are necessary at two or more time points to obtain accurate measures of change.

The termination criterion applied in a fully adaptive CAT will differ for each of these three testing applications. When the objective is to measure each examinee to the same degree of precision, CATs are typically ended when the observed standard error associated with the θ estimates reaches a predefined value. Thus, for example, a CAT can be ended when every examinee's SEM is less than or equal to 0.25. This would result in θ estimates that had 95% confidence intervals that spanned a half θ unit in either direction , e.g., a θ estimate of 1.0 would have a (95%) error band that ranged from 0.5 to 1.5. Because of variations in the information structure of an item bank at various levels of θ, and because of individual differences in the consistency with which examinees answer test items, obtaining such equiprecise measurements will require that the number of items administered to each examinee be allowed to vary.

Testing for classification uses a different CAT termination criterion. After a cutoff score has been expressed on the θ scale, and a desired level of classification accuracy is determined, each successive θ estimate in a CAT is bounded by the appropriate SEM error band. For example, if a 95% confidence classification is desired, the error band would be ± 2 standard errors. As each item is administered, θ is re-estimated and the new error band constructed around it. Testing continues until (after some prespecified minimum number of items) the error band for a θ estimate is entirely above the cutoff score or below it. When this occurs, a "high confidence" decision can be made from the CAT results, which will be equal to or better than the prespecified level of classification accuracy. Again, to obtain this objective, the number of items administered to each examinee must be permitted to vary.

Measuring individual change has been particularly troublesome in psychological measurement due to the unreliability of change scores (e.g., Cronbach & Furby, 1970) and floor and ceiling effects that occur with scores from conventional tests. Using CAT with a specialized termination criterion can result in measures of change that have properties that better capture change than do scores from conventional tests (Kim-Kang & Weiss, 2008). In this application, a CAT termination criterion can use the SEM bands from a Time 2 CAT compared to those from a Time 1 CAT obtained from the same examinee to determine if significant change has occurred. The Time 2 CAT can be ended when the two error bands no longer overlap, indicating that significant change has occurred, or when a sufficient number of items has been administered and it becomes clear that significant change has not occurred (Nydick & Weiss, 2010). Again, because of wide individual differences in test performance among examinees, combined with individual variations in magnitudes of change, CAT test length must be allowed to vary across examinees. Finkelman, Weiss, and Kim-Kang (2010) proposed and evaluated hypothesis testing methods for evaluating individual change and the accuracies of those methods using variable terminating CATs.

Putting It All Together: Examples of Fully Adaptive CATs

Equiprecise CAT. Figure 7 shows a sample CAT report from a CAT designed to measure each examinee to a pre-specified level of precision (minimum SEM of .20). To keep the test to a reasonable length, a maximum of 40 items was specified. In this particular test, the test was terminated when the 40-item maximum was reached.

The report is a graphic plot of the examinee's progress through the CAT after each item has been administered. A "C" for an item plotted at the current θ estimate indicates that the item was answered correctly; an "I" indicates that the item was answered incorrectly. The dashed lines represent a two SEM band around the θ estimate. The beginning θ for this CAT (represented by an "X" at $\theta = -0.24$) was based on a randomly selected θ in the range ±1.0. Rather than using a Bayesian θ estimate after the first item was answered, this CAT used an alternate method—the most difficult item in the bank was administered to attempt to force a mixed response pattern so that maximum likelihood estimation could be used. Since Item 2 was also answered correctly, the next most difficult item was administered, which was answered incorrectly.

As Figure 7 shows, generally, a correct answer is followed by an increase in the θ estimate and an incorrect answer is followed by a decrease in the θ estimate. The figure also shows the convergence in θ estimates—the differences between successive θ estimates are large at the beginning of the CAT and tend to become smaller as the CAT progresses. With a few exceptions, the SEM tends to decrease as each item is answered and the differences between successive θ estimates tend to decrease as more items are answered. The figure also shows that the CAT began to converge after about Item 10, with changes in θ estimates occurring in the first decimal place. Similar to the Binet adaptive test, the CAT selected the most appropriate range of items from the bank for this examinees—except for the first eight items, all items administered to this examinee were items that would be answered correctly about 50% of the time by examinees whose θs were between 0.9 and 1.75. More difficult items and easier items in the bank were not administered to this examinee.

Figure 8 shows the results of an equiprecise CAT for a different examinee. The entry θ estimate for this test was $\theta = 0.0$ and the first item was correctly answered. As a consequence, the second item was again the most difficult item in the bank, which was answered incorrectly resulting in a maximum likelihood θ estimate of 0.11 and an SEM = 0.52. This θ estimate was then used to select Item 3. The CAT response record shows a quick convergence of the θ estimates for this examinee accompanied by a rapid reduction in the SEMs. Had the test used a termination SEM of 0.20, the CAT could have been terminated after 17 items with a θ estimate that differed from the 30-item θ estimate in the second decimal place; an SEM termination value of 0.25 would have terminated the CAT after 9 items with a θ estimate of −0.25, which is very close to the 30-item θ of −0.21. Because the final θ estimate after the limit of 30 items was very close to the starting value of $\theta = 0.0$, a very narrow range of items was administered to this examinee from the larger CAT bank—with the exception of the second item, items administered were those appropriate for examinees with θs between 0.20 and −0.38. This response record also illustrates another feature of most CATS: Excluding the first few items in a CAT, the proportion correct for the majority of examinees will converge to $p = 0.50$. Excluding the first two items (which were not based on estimated θ), 15 of 28 items were correctly answered for a proportion of 0.54.

This test will terminate when the standard error of theta is equal to or less than 0.200
Minimum number of items = 5 Maximum number of items = 40
Theta was estimated by maximum likelihood.

Examinee Name : John Q. Public

The standard error band plotted as ---- is plus or minus 2.00 standard errors.
X = Initial theta value C = Correct answer I = Incorrect answer

```
Item  Theta  SE   -3.......-2........-1.........0........+1........+2........+3
  0  -0.24* 1.00*        --------------------X--------------------
  1   4.00* 1.00*                            .          ---------------------->
  2   4.00* 1.00*                            .          ---------------------->
  3   2.52  0.84                             .        ------------------I-----
  4   2.77  0.68                             .             -------------C---
  5   2.38  0.61                             .           ------------I-------
  6   2.09  0.61                             .        ------------I----------
  7   1.49  0.89                          ------------------I----------------
  8   0.36  1.00                 --------------------I--------------------
  9   0.88  0.63                          ------------C-------------
 10   1.13  0.56                             -----------C-----------
 11   1.34  0.49                             .  ----------C----------
 12   1.44  0.46                             .    ---------C---------
 13   1.55  0.43                             .     ---------C---------
 14   1.67  0.41                             .       --------C--------
 15   1.54  0.38                             .      --------I--------
 16   1.60  0.36                             .       --------C-------
 17   1.70  0.35                             .        -------C-------
 18   1.76  0.34                             .         -------C-------
 19   1.65  0.32                             .         ------I------
 20   1.52  0.31                             .       -------I------
 21   1.40  0.30                             .      -------I------
 22   1.27  0.30                             .     ------I------
 23   1.30  0.28                             .      ------C-----
 24   1.32  0.28                             .      ------C-----
 25   1.36  0.27                             .       -----C------
 26   1.40  0.27                             .       -----C------
 27   1.31  0.26                             .      ------I-----
 28   1.34  0.25                             .       -----C-----
 29   1.37  0.25                             .       -----C-----
 30   1.40  0.24                             .        ----C-----
 31   1.43  0.24                             .        -----C-----
 32   1.46  0.24                             .        -----C-----
 33   1.50  0.24                             .         ----C-----
 34   1.53  0.23                             .         -----C----
 35   1.55  0.23                             .         -----C-----
 36   1.59  0.23                             .          ----C-----
 37   1.62  0.23                             .          -----C----
 38   1.58  0.22                             .          ----I-----
 39   1.53  0.22                             .         -----I----
 40   1.55  0.22                             .          ----C----
```

*Arbitrarily assigned value.
The final theta estimate based on 40 items was 1.55 with a standard error of 0.22, resulting in a 2.00 standard error band of 1.11 to 1.99
This test was terminated when the maximum number of items was reached.

Figure 7. A Sample Report on an IRT-Based Adaptive Test

```
Item  Theta   SE   -3.......-2........-1.........0........+1........+2........+3
  0   0.00*  1.00*           --------------------X--------------------
  1   4.00*  1.00*                              .          --------------------->
  2   0.11   0.52                        -----------I----------
  3   0.20   0.45                          ----------C---------
  4  -0.04   0.35                          -------I-------
  5   0.05   0.32                            ------C------
  6  -0.13   0.29                          ------I------
  7  -0.07   0.27                           ------C-----
  8  -0.18   0.25                           -----I-----
  9  -0.25   0.25                          -----I-----
 10  -0.18   0.23                           -----C----
 11  -0.27   0.23                          -----I----
 12  -0.21   0.22                           ----C-----
 13  -0.26   0.22                           ----I----
 14  -0.34   0.22                          ----I-----
 15  -0.37   0.22                         -----I----
 16  -0.33   0.20                          ----C----
 17  -0.29   0.19                           ----C---
 18  -0.33   0.19                          ----I----
 19  -0.38   0.19                          ----I----
 20  -0.34   0.18                          ----C----
 21  -0.30   0.18                           ----C---
 22  -0.27   0.17                           ----C---
 23  -0.29   0.17                           ----I---
 24  -0.26   0.17                            ---C---
 25  -0.28   0.16                          ----I---
 26  -0.30   0.16                          ----I---
 27  -0.27   0.16                           ---C---
 28  -0.25   0.15                           ---C---
 29  -0.23   0.15                           ---C---
 30  -0.21   0.15                           ---C---
```

*Arbitrarily assigned value.

The final theta estimate based on 30 items was -0.21 with a standard error of 0.15, resulting in a 2.00 standard error band of -0.51 to 0.08

Figure 8. A Sample Report on an IRT-Based Adaptive Test for a Different Examinee

Classification CAT. Figure 9 shows a response record resulting from a CAT designed to make a dichotomous classification. For this purpose, the CAT was implemented similarly to those in Figures 7 and 8, except for the termination criterion. The test was designed to end when the SEM band surrounding a θ estimate fell below a prespecified θ cutoff value. The SEM error band in this case was ±1 SEM (resulting in a 68% two-tailed confidence interval) and the cutoff value was $\theta = +1.0$ (as indicated by the vertical dashed line in the figure). A minimum of 10 items was specified to avoid premature test termination and a maximum of 50 items was specified to avoid excessive testing times.

As with the CATs in Figures 7 and 8, the first item (based on a starting $\theta = 0.0$) was answered correctly and three of the most difficult items in the bank were given until an incorrect answer was obtained. Two incorrect answers then were followed by a string of responses essentially alternating between correct responses to less difficult items and incorrect responses to slightly more difficult items. As a consequence, the examinee's θ

This test terminated when the theta estimate plus or minus 1.00 standard errors
was above or below a theta cutoff of 1.00.
Minimum number of items = 10 Maximum number of items = 50
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 9. CAT Response Record for a Dichotomous Classification CAT

estimates slowly decreased from an estimated high of $\theta = 2.66$ to a low of 0.55 at Item 25. At Item 25, the θ estimate and the specified SEM band were completely below the cutoff value of $\theta = 1.0$, and the test was terminated. Note that the SEM value at termination was 0.43, which is fairly high, but it was not necessary to continue the test to reduce the SEM, since for classification purposes the test's termination criterion was met. The results show that the θ estimates were beginning to converge at around Item 21 and the SEMs began to display convergence (albeit at a high value) at Item 12. This response record also illustrates a phenomenon not evident in the other two response records: The SEMs in Figure 9 increased slightly at Items 17, 19, 21, 23, and 25, suggesting that the examinee was not entirely responding in accordance with the IRT model used to estimate θ. This result partially accounts for the relatively high SEM observed after 25 items.

Constrained CAT

The fully adaptive CATs illustrated above are unconstrained. That is, items are selected based only on maximum information at the current θ estimate at each stage of the CAT. In some applications of CAT, however, item selection has to be constrained by incorporating non-psychometric criteria into the item selection process. The major types of constraints applied include item exposure, content balancing, and "enemy" items.

Item exposure becomes an issue in CAT when tests are used to make decisions about individuals that have important consequences for those individuals. Thus, when tests are used to select individuals for entry into a college or university, for admission into special programs that might benefit the individual financially, for hiring into a particular job or position, or for licensure or certification, such "high-stakes" consequences sometimes motivate examinees to attempt to obtain information on test items so that they can enhance their scores. Because CAT testing programs tend to be continuous—tests are given to examinees over a long time period—examinees who have taken a CAT might remember some test items and make that information available to examinees who subsequently take the test. To minimize this potential problem, item selection based on item information can be constrained to (probabilistically) "expose" each item to some maximum proportion of examinees (see Georgiadou, Triantafilou, & Economides, 2007, for a comprehensive review of item exposure control methods). As a consequence, more items are used from a given item bank, but no items will be seen by all or a large number of examinees.

Some tests, although developed to meet the unidimensionality assumption required for the use of most IRT models to implement CAT, consist of items that vary in content characteristics. For example, a mathematics test used to measure math achievement in the early school grades might consist of items measuring addition, subtraction, multiplication, and division. Similarly, a depression scale might include items that reflect various aspects of depression (e.g. dysfunction in cognition, overt behavior, or mood, and somatic symptoms). In both cases, different item content might have different levels of difficulty, yet the scale is unidimensional. For certain applied purposes, it might be important to ensure that for a given examinee their CAT includes a proportionate sampling of items from each of the content domains. Thus, CATs can be constrained to provide (approximate) pre-defined proportions of items from content domains that comprise the CAT item bank. Kingsbury and Zara (1989, 1991) provide a review of some methods to achieve content balancing.

A third type of constraint frequently implemented in CAT is that of eliminating "enemy items." In some testing situations, some items in the bank provide clues that might be useful in answering other items; or some items might be very similar to other items (e.g., minor rewordings) so that their administration to a given examinee would be redundant, as well as violating the assumption of local independence that underlies IRT-based CAT—that the responses to test items are independent of each other except for their reliance on the trait that underlies the set of items. To control for enemy items, a CAT can include a list of subsets of items that should not be administered together. If an examinee answers any item in the subset, none of the other items in that set are administered to that examinee.

Unconstrained CATs will be the most efficient, so item selection constraints are generally used only as required in a particular CAT. Because any constraints imposed in a CAT will result in the selection of items that are suboptimal from a psychometric point of view (i.e., provide less information and, therefore, result in less rapid convergence of θ estimates), unless an item bank has many items that are replicates or near replicates of each other in terms of item information, constrained CATs will typically require the administration of more items to achieve the same degree of measurement precision or classification accuracy than unconstrained CATs.

Conclusions

Conventional peaked tests, developed using last century's methods of instrument construction, can measure well if—and only if—an examinee's level on a trait matches the region of the trait where the test is peaked. However, the purpose of measurement is to determine where an examinee's trait level is located on the trait, and it cannot be known in advance. As demonstrated above, as the examinee's trait level deviates from the test's location, measurement becomes extremely poor with very large errors of measurement. These errors of measurement result in conventional score variabilities that are artificially inflated by random error, reducing the utility of the scores for use in the most simple—as well as the most complex—statistical analyses. Error-laden standard deviations and variances will reduce the power of *t* tests or complex analyses of variance to detect differences in means and will similarly introduce error into all types of correlational analyses.

Computerized adaptive testing provides a viable solution to these problems. Because CATs are dynamic, adjusting the test to each examinee as the test is administered, they are both efficient and effective. CATs are effective because they essentially deliver a peaked test to each examinee; that is, they quickly adapt to the examinees's trait level as the test is being delivered to identify the subset of items in a pre-calibrated item bank that will best measure each examine. That subset of items is the subset on which the examinee will get about 50% of the items correct. Because fully adaptive CAT selects items by maximum item information at the current trait estimate, they will also be efficient—they will use a minimum number of items to measure each examinee to a minimum standard error of measurement or to a predetermined degree of precision required for a particular application. As demonstrated above, CATs function well (e.g., with equal precision) for examinees at all levels of a trait.

Developing a CAT is more complex than developing a conventional test. They require relatively large item banks that are calibrated with IRT, and because they require certain decisions to be made that interact with the structure of an item bank, require the use of software such as CATSim (Weiss & Guyer, 2010) for proper design prior to implementing them. Thompson and Weiss (2011) provide an overview of the steps necessary to develop a CAT. But in spite of the increased complexity, the better measurements provided by CAT, and the resulting more accurate and precise data, are very likely to result in more meaningful research conclusions (as well as better decisions made based on individual measurement data) than are error-laden measurements from "off-target" peaked conventional tests.

References

Allen, M. J. & Yen, W. M. (1979; reissued 2002). *Introduction to measurement theory*. Prospect Heights IL: Waveland Press.

Binet, A., & Simon, T. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11,* 191–244.

Chang, H.-H., Qian, J., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211–222.

Chang, H.-H. & van der Linden, W. J. (2003) Optimal stratification of item pools in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement, 27,* 262-274.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-234.

Cronbach, L.J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin, 74*, 68–80.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

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

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

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201-212.

Finkelman, M., Weiss, D.J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34, 238-254.*

Georgiadou, E., Triantafilou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment, 5(8)*. Available at www.jtla.org.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D., K., Stover, A., Bock, R. D., Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59(4),* 49-58.

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

Kang, S.-M. & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29,* 87-105.

Kim-Kang, G. & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift für Psychologie / Journal of Psychology*, *216,* 49–58.

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

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.

Nydick, S. J. & Weiss, D. J. (2010). Accepting the null: Determining no change within the adaptive measurement of change. Paper presented at the 2010 International Association for Computerized Adaptive Testing conference. Arnhem, The Netherlands.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests *Practical Assessment, Research, and Evaluation, 16(1).*

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

Weiss, D. J. & Guyer, R. (2010). *Manual for CATSIM: Comprehensive simulation of computerized adaptive testing*. St. Paul MN: Assessment Systems Corporation.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. Chapter 18 in W. J. van der Linden and C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 355-372.). New York: Springer.