Encyclopedia of Social Measurement, Academic Press

Computerized Adaptive Testing

Daniel O. Segall

Defense Manpower Data Center
United States Department of Defense

Outline

I. CAT Response Models
II. Test Score Precision and Efficient Item Selection
III. Maximum Likelihood Approach
IV. Bayesian Approach
V. Item Selection Enhancements
VI. Item Pool Development
VII. Trends in Computerized Adaptive Testing

Glossary

content balancing a set of one or more ancillary item-selection constraints based on content or non-statistical item features.

conventional testing an approach to individual difference assessment where all examinees receive the same items, typically (but not necessarily) in printed mode.

exposure control algorithm an algorithmic enhancement to precision-based item selection that limits the usage rates of some highly informative items for the purpose of increased test security.

information a statistical concept related to the asymptotic variance of maximum likelihood trait estimates; it can be expressed as the sum of individual item information functions which can be evaluated at specific points along the trait scale.

item pool a collection of test questions and associated item-parameters from which items are selected for administration by the adaptive item-selection algorithm.

item response function a mathematical function providing the probability of a correct response conditional on the latent trait level θ.

measurement precision an index of the accuracy of test scores, often assessed by the average or expected squared difference between true and estimated trait parameters, $\mathrm{E}(\theta - \hat{\theta})^2$.

measurement efficiency the ratio of measurement precision to test-length: One test or testing algorithm is said to be more efficient than the other if it provides more precise scores for a fixed test-length, or if it achieves equally precise scores with fewer administered items.

stopping rule the rule used to determine when to end the test; typically based on the number of administered items (fixed-length), or on the precision-level of the estimated trait parameter (variable-length).

trait a psychological dimension of individual differences that includes ability, aptitude, proficiency, attitude, or personality characteristics.

trait estimate an item-response-theory based test score, denoted by $\hat{\theta}$, typically calculated by Bayesian or maximum likelihood estimation approaches.

trait parameter an item-response-theory based parameter θ that denotes the examinee's standing along the latent trait dimension.

COMPUTERIZED ADAPTIVE TESTING is an approach to individual difference assessment that tailors the administration of test questions to the trait level of the examinee. The computer chooses and displays the questions, and then records and processes the examinee's answers. Item selection is adaptive—it is dependent in part on the examinee's answers to previously administered questions, and in part on the specific statistical qualities of administered and candidate items. Compared to conventional testing where all examinees receive the same items, computerized adaptive testing (CAT) administers a larger percentage of items with appropriate difficulty levels. The adaptive item selection process of CAT results in higher levels of test-score precision and shorter test-lengths.

I. CAT Response Models

Modern CAT algorithms are based on concepts taken from item response theory (IRT) and from maximum likelihood and Bayesian statistical estimation theories. Early pioneers of CAT, including Frederic M. Lord and David J. Weiss (upon whose work modern CAT algorithms are based) used item response functions (IRFs) as the basic building blocks of CAT. These functions, denoted by $P_i(\theta)$, express the probability of a correct response for an item as a function of latent trait level θ. The trait estimated from adaptive testing can be a psychological (or other) dimension of individual differences, including ability, aptitude, proficiency, attitude, and personality. For ability measurement, IRFs are generally assumed to be monotonically increasing functions. Consequently, as θ increases, so too does the probability of a correct response.

One of the most commonly used mathematical expressions for an IRF is the three parameter logistic (3PL) model:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \ , \tag{1}$$

where the parameters a_i, b_i, and c_i denote the slope, difficulty, and guessing parameters, respectively for item i. The 3PL is often used to model dichotomously scored responses from multiple choice items. The two parameter logistic (2PL) model (often used to model attitude or personality items) is a special case of (1), where guessing is assumed to be nonexistent (i.e., $c_i = 0$). The one parameter logistic (1PL) model (where $a_i = 1$ and $c_i = 0$) is used in cases where the IRF associated with item i is characterized by its difficulty parameter b_i; all IRFs have identical slopes, and the probability of an examinee with infinitely low trait-level correctly answering the item is zero. Other IRF models have also been used to extract information from incorrect options of multiple choice items, or from other item response formats (e.g., rating scales).

According to the assumption of local independence, the conditional probability of an observed response pattern is given by the product of item specific terms:

$$P(u_1, u_2, ..., u_n|\theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \ , \tag{2}$$

where u_i denotes the scored response to item i ($u_i = 1$ if item i is answered correctly; $u_i = 0$, otherwise), $Q_i(\theta) = 1 - P_i(\theta)$ (i.e., denotes the conditional probability of an incorrect response) and n denotes the number of answered questions. One implication of (2) is that

the probability of a correct response to item i is independent of the response to item j after controlling for the effects of θ.

Another important property of IRT is *scale invariance*: The scale of measurement along which examinees are placed, the θ-scale, is defined independently of the statistical properties of the administered items. This invariance property does not hold for scales derived from classical test theory, which are founded on number or percentage correct scores. A percent-correct score of 75 on a test containing easy items has a different meaning than a score of 75 on a test containing difficult items. In contrast, an IRT based test-score (i.e., trait estimate $\hat{\theta}$) has the same meaning for tests containing either easy or difficult items (provided all item parameters have been transformed to a common scale). This IRT invariance property enables the comparison of scores from different or overlapping item-sets. In the context of IRT, $\hat{\theta}$ test-scores are all on a common measurement scale, even though these scores might have been estimated from tests consisting of different items.

II. Test Score Precision and Efficient Item Selection

Although the invariance property of IRT ensures that the interpretation of θ remains constant across tests consisting of different items, the precision with which θ can be estimated is very much dependent on the statistical properties of the administered items. Examinees with high θ-levels can be most accurately measured by tests containing many difficult items; examinees with low θ-levels can be most precisely measured by tests containing many easy items. This can be verified, for example, by an examination of the 1PL model, where the asymptotic variance of the maximum likelihood estimator is given by

$$\mathrm{Var}(\hat{\theta}|\theta) = \left[1.7^2 \sum_{i=1}^{n} P_i(\theta)Q_i(\theta)\right]^{-1} . \tag{3}$$

It can be seen from (3) that the smallest variance is obtained when $P_i(\theta) = Q_i(\theta) = 1/2$ for each item—any other values of these conditional response probabilities lead to a larger variance. From (1), we see that for the 1PL this optimal condition occurs when $b_i = \theta$, that is when the difficulty parameter of each item matches the examinee trait-level parameter.

One implication of (3) is that the optimal (i.e., most precise) testing strategy chooses items solely on the basis of the examinee's true trait-level θ. But obviously this is not possible, since θ is unknown prior to testing. (If it were known, testing would be unnecessary in the first place.) It is possible however to use an iterative adaptive algorithm, where an estimated trait-level $\hat{\theta}_k$ is obtained after each administered item $k = 1, ..., n$, and the difficulty parameter of the next administered item b_{k+1} is matched to the current estimate: $b_{k+1} = \hat{\theta}_k$. In this sense, the difficulty of the next question b_{k+1} is adapted to the most up-to-date trait estimate $\hat{\theta}_k$ of the examinee. By doing so, the precision level of the final estimate (obtained after the completion of the last item) is greater than that expected from conventional non-adaptive testing.

This idea of adapting the statistical properties of administered items based on responses to previous items forms the basis of all CAT item selection algorithms. However, commonly used algorithms differ along two primary dimensions: first in the type of statistical estimation procedure used (maximum likelihood versus Bayesian), and second in the type of item-response model employed (e.g., 1PL, 2PL, or 3PL).

III. Maximum Likelihood Approach

The maximum likelihood (ML) approach to CAT item-selection and scoring is based on the log-likelihood function

$$l(\theta) = \ln \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \ . \tag{4}$$

The estimate $\hat{\theta}_{(\mathrm{ML})}$ is defined as the value of θ for which the likelihood (or equivalently the log-likelihood) function is maximized. Since no closed-form expression exists for $\hat{\theta}_{(\mathrm{ML})}$, it is typically calculated using an iterative numerical procedure such as the Newton-Raphson algorithm.

The estimator $\hat{\theta}_{(\mathrm{ML})}$ is asymptotically normally distributed with mean θ and variance

$$\begin{aligned} \mathrm{Var}(\hat{\theta}|\theta) &= \left[-\,\mathrm{E}\,\frac{\partial^2}{\partial\theta^2} l(\theta) \right]^{-1} \\ &= 1/\sum_{i=1}^{n} I_i(\theta) \ , \end{aligned} \tag{5}$$

where the information function for item i, denoted by $I_i(\theta)$, is

$$I_i(\theta) = \frac{\left[P_i'(\theta) \right]^2}{P_i(\theta) Q_i(\theta)} \ , \tag{6}$$

and where $P_i'(\theta)$ denotes the derivative of the item response function with respect to θ. For the one and three parameter logistic models, these derivatives are $P_i'(\theta) = 1.7 P_i(\theta) Q_i(\theta)$ and $P_i'(\theta) = 1.7 a_i Q_i(\theta) \left[P_i(\theta) - c_i \right] / (1 - c_i)$, respectively.

From (5), it is clear that the asymptotic variance of the ML estimate $\hat{\theta}_{(\mathrm{ML})}$ can be minimized by choosing items with the largest information values. If θ were known in advance of testing, then available items could be rank-ordered in terms of their information values (6) at θ, and the most informative items could be selected and administered. Since θ is not known (to know or approximate θ is of course the purpose of testing), the most informative item can be selected using item information functions evaluated at the provisional (most up-to-date) trait estimate, $I_i(\hat{\theta}_{k(\mathrm{ML})})$. After the chosen item has been administered, and the response scored, a new provisional estimate can be obtained and used to reevaluate item information for the remaining candidate items. These alternating steps of trait estimation and item-selection are repeated until a stopping rule (typically based on test-length or precision) is satisfied. The adaptive item selection and scoring algorithm is summarized in Table 1.

IV. Bayesian Approach

In instances where a prior distribution for θ can be specified, some test developers have opted to use a Bayesian framework for item-selection and trait estimation. The prior density, denoted by $f(\theta)$, characterizes what is known about θ prior to testing. The most common

Table 1: CAT Item Selection and Scoring Algorithm

Step		Description
1.	Calculate provisional trait estimate.	Obtain a provisional trait estimate, $\hat{\theta}_k$, based on the first k responses.
2.	Choose Item.	Compute information $I_i(\hat{\theta}_k)$ for each candidate item by substituting the provisional trait estimate $\hat{\theta}_k$ (calculated in Step 1) for the true parameter θ in (6). Select for administration the item with the largest item information value.
3.	Administer item and record response.	
4.	Repeat Steps 1 – 3	until the stopping rule has been satisfied.
5.	Calculate final trait estimate $\hat{\theta}$	based on all responses, including the response to the last administered item.

approach to prior-specification in the context of CAT sets the prior equal to an estimated θ-density calculated from existing (or historical) examinee data. Then the assumption is made that future examinees (taking the CAT test) are independent and identically distributed $\theta \overset{iid}{\sim} f(\theta)$. Although in many cases, additional background information is known about examinees relating to θ (such as subgroup membership), this information is often ignored in the specification of individual examinee priors—to allow such information to influence the prior could lead to, or magnify, subgroup differences in test-score distributions.

A Bayesian approach provides estimates with different statistical properties than provided by ML estimates. In CAT, Bayesian estimates tend to have the advantage of smaller conditional standard errors $\sigma(\hat{\theta}|\theta)$, but possess the disadvantage of larger conditional bias $B(\theta) = \mu(\hat{\theta}|\theta) - \theta$, especially for extreme θ levels. Thus the choice of estimation approach involves a trade-off between small variance (of Bayesian estimates) and small bias (of ML estimates). Bayesian procedures do in general provide smaller mean-squared-errors (MSEs) between θ and $\hat{\theta}$ (which is a function of both conditional variance and bias) than provided by ML estimates. This suggests that Bayesian estimates can provide higher correlations with external criteria, and a more precise rank-ordering of examinees along the θ-scale. Practitioners who are concerned about the effects of bias, or who do not have precise estimates of the trait distribution tend to favor the ML approach. Conversely, practitioners whose primary objective is to minimize MSE or conditional variance have tended to favor Bayesian approaches.

The Bayesian approach to CAT item-selection and scoring is based on the posterior

density function

$$f(\theta|u) \propto f(u|\theta)f(\theta) \, , \tag{7}$$

where $f(u|\theta)$ is equivalent to the probability function (2), $f(\theta)$ is the prior distribution of θ, and $u = (u_1, ..., u_n)$ is a vector of scored responses. Whereas the prior $f(\theta)$ describes what is known about θ before the data are observed, the posterior density function $f(\theta|u)$ provides a description of what is known about the examinee's trait-level after the item response data u have been obtained. Typically, summary statistics are used to characterize the posterior distribution: a measure of central tendency (such as the posterior mean or mode) is often taken as the trait point estimate, and the variance of the posterior distribution is typically taken as a measure of uncertainty. Small posterior variance values suggest that $\hat{\theta}$ has been estimated with a high degree of precision; large posterior variance values suggest otherwise.

One Bayesian approach to item selection chooses the next item to minimize the expected posterior variance, where the expectation is taken with respect to the yet-to-be observed response to the candidate item. This quantity is calculated by computing the values of the posterior variance if the candidate item is answered both correctly and incorrectly, and then calculating a weighted average of the two posterior variances, where the weights are equal to the probability of correct and incorrect responses based on the predictive posterior distribution.

A less computationally intensive and more commonly used Bayesian item-selection method is consistent with a normal-based inference approach. According to this approach, the posterior distribution is approximated by a normal density

$$f(\theta|u) = \mathrm{N}(\hat{\theta}_{(\mathrm{MAP})}, V) \, , \tag{8}$$

with mean equal to the mode (maximum *a posteriori;* [MAP]) of the posterior density, denoted by $\hat{\theta}_{(\mathrm{MAP})}$, and variance based on the expected information evaluated at the mode:

$$\begin{aligned} V &= 1/\left\{ -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \ln f(\theta|u) \right] \right\}_{\theta=\hat{\theta}_{(\mathrm{MAP})}} \\ &= 1/\left\{ 1/\sigma^2 + \sum_{i=1}^{n} I_i(\theta) \right\}_{\theta=\hat{\theta}_{(\mathrm{MAP})}} \, . \end{aligned} \tag{9}$$

This approximation assumes that the prior is normal with variance denoted by σ^2. The information function for item i, denoted by $I_i(\theta)$ is equivalent to the one derived for the ML case given by (6). It is clear from an examination of (9) that the administration of the item with the largest information value (evaluated at $\hat{\theta}_{k(\mathrm{MAP})}$) will provide the greatest reduction in posterior variance V. As with the ML approach, the adaptive item selection and scoring algorithm summarized in Table 1 is used, where the provisional trait estimate $\hat{\theta}_k$ is set equal to the posterior mode $\hat{\theta}_{k(\mathrm{MAP})}$. Calculation of the mode requires the use of an iterative numerical algorithm to find the maximum of the log posterior density function (7). Alternating steps of trait estimation and item-selection are repeated until a stopping rule is satisfied. The posterior variance based on observed information

$$\mathrm{Var}(\theta|u) = 1/\left[-\frac{\partial^2}{\partial\theta^2} \ln f(\theta|u) \right]_{\theta=\hat{\theta}_{(\mathrm{MAP})}} \tag{10}$$

is an often-used characterization of measurement precision.

V. Item Selection Enhancements

Although the adaptive item-selection algorithms form an efficient basis for precise measurement, test developers have often found it beneficial or necessary to alter these algorithms. These alterations, or enhancements, include the specification of rules used to choose the first several items; the specification of rules used to stop the test; modifications to the item-selection algorithms intended to reduce opportunities for test-compromise, and to help achieve a more balanced item content; and the use of time-limits.

A. Item-Choice Early in the Adaptive Sequence

Most commonly used adaptive item-selection algorithms require the existence of a provisional trait estimate. This provisional estimate is used to evaluate the relative information contribution of candidate items, and is specified from the responses to earlier items. But how should the first item be selected? The choice of the first item and other early items depends on the approach taken: ML or Bayesian.

ML approaches have adopted a set of heuristics for item selection early in the adaptive sequence. Typically, the first item selected is one of moderate difficulty relative to the population of examinees. If the first item is answered correctly, then a more difficult item is selected and administered; if the first item is answered incorrectly then an easier item is selected. If necessary, selected items become successively easier or harder until at least one correct and incorrect response has been obtained. At this point, the ML function will typically possess a finite maximum, and the adaptive item selection and scoring algorithm (Table 1) can be used.

The Bayesian approach formalizes these heuristics by setting the initial provisional trait estimate equal to the mean of an informative prior trait density. The first item chosen is one with high—or highest information at the prior mean. After the administration of the first item, the provisional trait estimation and item selection algorithm (given in Table 1) can be applied in a straightforward manner. Unlike the ML estimate, the provisional Bayesian estimate (taken as the posterior mean or mode) is defined for all response patterns, including those containing all correct or incorrect responses.

B. Stopping Rules

There are two common test termination or stopping rules used in CAT: *fixed-length* and *variable-length*. Fixed-length tests require that the same number of items be administered to each examinee. One consequence of fixed-length tests is that measurement precision is likely to vary among examinees. In contrast, variable-length tests continue the administration of items until an individualized index of precision satisfies a target precision level. These precision indices are often based on ML information (5) or Bayesian posterior variance (10) statistics.

Test developers have found that the choice of stopping rule is often highly dependent on the test-purpose, item-pool characteristics, and operational constraints. In many instances for example, equally precise scores among examinees are paramount, helping to ensure that decisions and interpretations made on the basis of test-scores are equally precise for all examinees. In other instances however, the occasionally long test-lengths (possible with variable-length tests) might be judged too burdensome for examinees, and possibly for

test-administrators as well. To moderate some of the operational burdens, variable-length testing has been implemented with upper-bound constraints on the maximum number of administered items, and in some instances, on the maximum amount of testing time allowed for each examinee. In other instances, test developers have opted for fixed-length tests to help standardized testing-conditions including variability in testing-time, and related testing-fatigue.

C. Test Compromise Safeguards

In some instances, examinees may attempt to misrepresent their performance. This is especially likely when the test-scores on the exam are used as a basis for important decisions. With CAT, the same items are typically administered on multiple occasions (spanning weeks, months, or possibly years). This repeated item use provides examinees with an opportunity to obtain information about the questions from others taking the test before them. In these instances, one or more compromise deterrents can be implemented.

The adaptive item selection algorithm (Table 1) provides highly efficient, but deterministic item selection. Consequently, two examinees providing the same pattern of responses to a set of multiple choice questions (e.g., A, D, C, A, B, C, D, ...) will receive the same items, and the same θ estimate. An examinee could be assured a high score by simply re-entering the response pattern copied from a high-scoring examinee. This strategy can be thwarted however by adding a stochastic component to the item selection algorithm. Rather than matching the difficulty parameter b_i with the provisional trait estimate $\hat{\theta}_k$, the next administered item for the 1PL model can be selected at random from among those items with difficulty parameters b_i falling in a narrow interval around $\hat{\theta}_k$ (namely the interval $\hat{\theta} \pm \delta$, where δ is some suitably small constant). A strategy with similar intent designed for the 3PL model is based on the 5-4-3-2-1 algorithm, where the first item is selected at random from the five most informative items at the current provisional trait level, the second administered item is selected at random from the four most informative items evaluated at the current provisional trait estimate, and so forth. The fifth and subsequent items are chosen to maximize precision evaluated at the provisional trait estimate.

Although these strategies decrease or eliminate the gains associated with copied answer patterns, they do not necessarily limit the usage or exposure of the item pool's most informative items. That is, these strategies can still lead to instances where some items are administered to nearly all examinees. An alternate method, referred to as the Sympson-Hetter exposure control algorithm was designed specifically to place an upper ceiling on the administration rates of the most used items (typically highly discriminating items of moderate difficulty).

The Sympson-Hetter exposure control algorithm assigns an exposure control parameter, denoted by e_i, to each item i. These parameters are used in conjunction with the adaptive item selection algorithm to screen items. For the selection of the kth item, candidate items are rank-ordered by information level evaluated at the provisional trait estimate $\hat{\theta}_k$. The item with the largest information is considered first. A random uniform number r (between 0 and 1) is drawn; the item either passes or fails the exposure screen: If $r \leq e_i$ then item i passes and is administered; otherwise it fails and is not considered again for administration to the examinee. If the first evaluated item fails the exposure screen, then the next most informative item is considered for administration. A new random number

is drawn, and the exposure screen is repeated. This screening process is repeated until a candidate item passes.

The exposure control parameters e_i are specified prior to testing, and are calculated through a series of computer simulations. The assigned e_i values are dependent on a target ceiling exposure value T, and on an assumed trait distribution $f(\theta)$. The use of the exposure control parameters ensures (in expectation) that the exposure rates of the most used items will not exceed the target ceiling rate T in a population with trait distribution $f(\theta)$. In practice, the target ceiling exposure rate T is often set to a value between 1/10 and 1/3, ensuring that the most used items are not administered to more than 1/10 or 1/3 of the examinee population.

A conditional version of the Sympson-Hetter approach has been suggested for use in situations where it is important to maintain a target ceiling exposure rate for homogenous subpopulations of examinees. Over narrow ranges of θ, the unconditional approach can provide higher than desired exposure rates for some items, higher than the target T specified for the overall population. The conditional approach remedies this problem by using a vector of exposure parameters for each item $(e_{i1}, e_{i2}, ...)$, where the exposure parameter used e_{ij} is specific to both the item i, and to a narrow trait-range indexed by the subscript j. This trait-range is associated with the value of the provisional trait estimate. The conditional approach helps ensure that the exposure rates of the most used items do not exceed the target ceiling rate T; this assurance is made without requiring any specific assumption regarding the form of the trait distribution $f(\theta)$.

Other methods intended to further reduce item exposure are commonly used in conjunction with the Sympson-Hetter exposure control method. Two such methods include the simultaneous and sequential use of multiple item pools. In the case of simultaneous item pool use, examinees are randomly assigned to two or more distinct (non-overlapping) item pools. These item pools serve the same function as alternate test forms in conventional testing. In the case of sequential item pool use, the item pool is continuously updated or replaced over a period of days, weeks, or months, thus making sharing item content among examinees less profitable.

Inevitably, the choice of any CAT exposure control method requires a consideration of the effects on measurement efficiency and test-development costs. (*Measurement efficiency* is defined as the ratio of test-score precision to test-length.) In general lower maximum item exposure rates result in either lower measurement efficiency, or in higher test development costs associated with larger or more numerous item pools. When making decisions about exposure control algorithms, including decisions about target maximum exposure rates and the number or size of simultaneous or sequential item pools, test developers have considered the unique compromise pressure placed on their exams. As part of the evaluation process, test developers typically perform extensive simulation analyses to examine the consequences of exposure control algorithms on measurement precision (for fixed-length tests), and on test-lengths (for variable-length tests). These considerations have led some high-stakes developers to set low target maximum exposure levels at 0.10 and frequent item-pool replacement schedules (of just several weeks), and have led other test developers to use somewhat higher targets of 1/3 in conjunction with two or three simultaneous pools replaced at five-year intervals.

D. Content Balancing

Test developers have been compelled in many cases to depart from strict precision considerations when designing and implementing CAT item selection algorithms. These include cases, where for example, the item pool consists of items drawn from different content areas of a more general domain (e.g., math items drawn from algebra and geometry). In such instances, item selection algorithms which maximize precision may not administer properly balanced tests, resulting in test scores which have questionable validity. To help ensure adequately balanced content across examinees, constraints can be placed on the adaptive item selection algorithms (e.g., constraints that ensure equal numbers of administered algebra and geometry items).

The most basic approach to content balancing spirals the sequence of item administration among key content areas. For example, math items would be administered in the order: (1) algebra, (2) geometry, (3) algebra, (4) geometry, and so forth, where each item represents the most informative item (passing the exposure-control screen if used) at the provisional trait level among items in the given content (i.e., algebra or geometry) domain.

Although the spiraling approach is adequate for a small number of mutually exclusive content areas, this approach is poorly suited for situations where more complex content constraints are desired. Consider the case where for example, items are classified along several dimensions simultaneously, and as a result do not fall into mutually exclusive categories. In such cases, methods such as the Weighted Deviations or Shadow Testing approaches can be used. These approaches are designed to maximize precision while attempting (in the case of the former method), or forcing (in the case of the latter method) adherence to specified content constraints.

Test developers have placed different levels of emphasis on the issue of content balancing. Developers of licensure and certification exams for example have tended to produce CAT exams where content targets and constrains heavily influence item choice. In these exams, the direct demonstration of the understanding of key facts and concepts is considered so important that it is not sufficient to infer mastery of one concept from the correct answers to items assessing more difficult concepts. In some instances, the balanced administration of items is so important that test developers have opted for a testlet based approach, where balanced groups or sets of items are selected and administered. Within each group, items are balanced for content and span a relatively narrow range of difficulty. Thus in the testlet based approach, the difficulty level of the testlet item-group (rather than the individual item) is tailored to the level of the examinee.

The issue of content balancing is complicated not only by the question of when to balance, but by the question of how finely to balance. Inevitably more detailed balancing constraints will lower the measurement efficiency of the adaptive testing algorithm, with some constraints having larger effects than others. For example, in a fixed-length test of high school math skills, the forced administration of a large number of calculus items (that happened to be difficult because of their advanced content) would degrade precision over the middle and lower proficiency ranges. Examinees in these ranges would be better measured by the administration of items of more appropriate difficulty levels, such as those taken from introductory or intermediate algebra. This example illustrates that the imposition of some content constraints may actually lead to a significant decrease in precision or measurement

Table 2: Factors Affecting CAT Measurement Efficiency

Item Pool Characteristics	Algorithm Characteristics
1. Size	1. Stopping Rule
2. Item Parameter Distributions	2. Content Constraints
3. Content Coverage	3. Exposure Control

efficiency. Unfortunately, there are no universally accepted rules regarding the optimal balance between content and measurement efficiency considerations in the construction of item-selection algorithms. Rather, test developers routinely weigh these trade-off considerations in the context of the specific exam and its intended purpose to arrive at a suitable approach to content balancing.

E. Time Limits

The imposition of time-limits can in some instances significantly degrade CAT measurement precision, since the effects of time-pressure are not explicitly modeled by standard item-selection and scoring algorithms. Even in spite of this undesirable consequence, most high-stakes high-volume testing programs have implemented overall test time-limits for a number of reasons, including the desire to help reduce excessive test times. In instances where time-limits have been imposed, most test developers have chosen to implement long time-limits which provide most or nearly all examinees with an opportunity to answer all items without feeling rushed.

Although time-limits might be desirable from an administrative standpoint, their use raises opportunities for gaming and test-compromise in high-stakes testing. Low ability examinees would be well advised to answer as few items as allowed. Under ML scoring, these low-ability examinees could capitalize on measurement error, which is greatest for short tests. Under Bayesian scoring, these same low-ability examinees could capitalize on the positive bias introduced by the prior, which is also greatest for short tests. To help discourage such test-taking strategies associated with time-limits, test developers have implemented various scoring penalties applied to incomplete fixed-length tests. For variable-length tests, fair and equitable provisions must be made to help ensure that those requiring longer tests (to achieve the target precision level) are given sufficient time.

VI. Item Pool Development

Characteristics of the item pool (including size, item parameter distributions, and content coverage) directly impact CAT measurement efficiency and test-score validity. Furthermore, particular characteristics of the adaptive algorithm (such as the stopping rule, number and type of content balancing constraints, and type and level of exposure control) can interact with key item-pool characteristics to further affect measurement efficiency and test-score validity. These characteristics are listed in Table 2.

Large item pools are desirable from several standpoints. First, large item pools tend to contain a larger set of highly discriminating items which in turn can provide greater measurement efficiency: greater precision for fixed-length tests, and shorter test-lengths

for variable-length tests. Second, large pools are more likely to satisfy content balancing constraints, or satisfy them without severely impacting efficiency. For fixed-length tests, large pools enable lower exposure levels (for the most used items), and can satisfy these levels without severely impacting precision. Many test developers have found high precision levels can be obtained with pools whose size is about six to eight times the test length.

In principle, the ideal item pool contains items with difficulty parameters (b_i's) uniformly distributed throughout the θ range, and for the 3PL model contains high discrimination parameters (a_i's) and low guessing parameters (c_i's). In practice, these ideal parameter distributions are often difficult to achieve. For some tests, highly discriminating items may be rare, or may only exist for items with difficulty values that span a narrow range, or for items of specific content areas. In these cases, CAT algorithms can be very inefficient, resulting in test scores that have low precision over some trait ranges (for fixed-length tests), or resulting in long test-lengths (for variable-length tests). Consequently, test developers when possible, have tended to write and pre-test large numbers of items in hopes of ending up with a sufficient number of highly discriminating items of appropriate difficulty and content.

Standard CAT item selection and scoring algorithms assume that the IRFs for all items are known in advance. In practice these are estimated from examinee response data. For the 3PL model, large-scale testing programs have tended to use samples containing 500 or more responses per item to estimate item parameters. Programs that have based their item selection and scoring algorithms on the 1PL model have typically relied on smaller sample sizes for IRF estimation. Test developers routinely use conditional (on $\hat{\theta}$) item-score regressions to check model-fit. This model-fit analysis typically includes an additional check of dimensionality or local independence assumptions.

Many test developers have found it convenient, especially when developing the first set of pools to collect calibration data in paper-and-pencil format, since this mode of data collection is often faster and cheaper than collecting the same data by computer. In these cases, test developers have attempted to ensure that the use of item-parameter estimates obtained from paper-and-pencil data are adequate for use when the items are administered on computer in adaptive format. This assurance has been provided by several studies which have found inconsequential differences in item response functioning due to mode of administration (computer versus paper-and-pencil).

Because of the complexity of the interactions between item pool characteristics and adaptive testing algorithms, and the effects these have on measurement efficiency, test developers routinely conduct computer simulation studies to fine-tune the adaptive algorithms and to examine the adequacy of candidate item pools. These simulations take as input the item parameter estimates of items contained in the pool (a's, b's, c's), and if content balancing is proposed, the content classification of each item. Then, the consequences (on precision or test-length) of using the proposed adaptive testing algorithm can be examined for examinees falling at different trait levels. The output of these simulations are conditional (on θ) means and variances of the estimated scores $\hat{\theta}$. These simulation studies allow the effects of different variations of the adaptive algorithms (i.e., changes in content constraints, pool size, stopping rule, target exposure level, etc.) to be examined and compared. The outcome of these simulations are often used as a basis for determining the suitability of candidate item-pools and adaptive algorithms.

VII. Trends in Computerized Adaptive Testing

In recent years, research on item selection and scoring algorithms has continued. This includes work on item-selection algorithms intended to provide greater measurement precision. One class of approaches address the uncertainty regarding the provisional trait estimates toward the beginning of the test. These include such methods as the global information criterion, weighted likelihood information criterion, *a*-stratified method, and fully Bayesian approaches. Research has also continued on improved exposure control algorithms to further guard against test compromise. Another class of item-selection approaches have been developed to further increase the measurement efficiency of CAT in the context of multidimensional IRT modeling, where items are selected to maximize the information along several dimensions simultaneously.

As more testing programs have considered the use of CAT, more attention has been given to its cost-effectiveness. In addition to the benefits of increased measurement precision and reduced test lengths, CAT offers a host of other benefits associated with the computerized administration of test items. These include: immediate and accurate scoring, minimal proctor intervention, individually timed and paced test administration, standardized instructions and test administration conditions, improved physical test security (no hard-copy of test booklets are available for compromise), and provisions for handicapped examinees (large print, audio, and alternate input devices). Many of these benefits, especially when considered along side the key benefit of increased measurement efficiency, provide compelling incentives in favor of CAT. But several obstacles have prevented many test developers from adopting CAT. In addition to specialized software requirements (necessary for test development and administration), CAT also requires considerable resources for item pool development, and for the purchase and maintenance of computer test-delivery systems.

Compared to conventional testing paradigms, many high-stakes test developers have found that CAT requires more test items and greater data demands (for item calibration). These greater demands for items and data are due in part to requirements of the adaptive branching strategy, and in part to the change in testing schedule associated with the test delivery. Conventional exams administered in high or moderate stakes settings are often associated with periodic test schedules, where substantially different items (i.e., test forms) are administered on each testing occasion to help reduce instances of cheating. Because of the large item-pool development costs and the impracticality of administering a large number of computer-delivered exams on the same day (as is the case with periodic exams), CAT exams are administered exclusively using on-demand or continuous testing schedules. According to these schedules, the same items are used over an extended time period. These continuous schedules by their nature increase the opportunity for test compromise, which is most effectively countered by large numbers of items—either contained in a few large pools, or contained in a large number of smaller item pools.

It is noteworthy that increased demands for items and calibration data has been minimal for at least one major high-stakes testing program that transitioned from paper-and-pencil to CAT. This program (the Armed Services Vocational Aptitude Battery [ASVAB] program) differed from most programs in that its conventional test-version was also given on-demand, where compromise was controlled though the simultaneous use of a large num-

ber of alternate paper-based test-forms. It was found that a sufficient level of CAT precision and security could be achieved by the use of a small number of moderate sized item pools. The ASVAB experience suggests that a large part of the increase item/data demands typically associated with high-stakes CAT may be due to the change in testing schedule (from periodic to continuous), rather than to the demands of the adaptive item selection algorithms.

Research has also intensified on item pool data collection and maintenance procedures. Response data required for the calibration of new items can be easily obtained by administering these items along with the operational adaptive items. This method of seeding or interspersing helps ensure that examinees will provide high-quality motivated responses, and that the item parameter estimates obtained from these response data are appropriately scaled. Depending on the mixture of tryout and operational items presented to examinees, this sort of data collection design can raise special challenges for conventional item calibration approaches. Research has also continued on approaches for phasing items in and out of item-pools, and the effects of these approaches on precision and other important item-pool qualities.

Developers have been sensitive to computer literacy levels among their test-taking populations, and in particular to the literacy levels of particular (possibly economically disadvantaged) subgroups. In large-scale test development efforts, care has been taken to ensure that aspects of the computerized test-taking experience do not place particular subgroups at an unfair disadvantage relative to corresponding subgroups taking paper-and-pencil versions. Although no consistent subgroup/medium interactions along racial or gender lines have been identified, attempts have been made none-the-less to mitigate any possible disadvantage among subgroup members by the use of simple item presentation formats and by clear test-taking instructions. As computers become even more commonplace and ubiquitous among widely diverse segments of the population, this concern is likely to dissipate.

With the proliferation of computers in recent years, it would appear that one of the last major obstacles to CAT has been removed: the availability of computer platforms for CAT test delivery. However, many high-stakes test developers have been concerned about context effects associated with different hardware, and testing environments, and possible interaction effects of these with test-delivery software. For paper-and-pencil tests, test-performance can be affected by subtle differences in booklet font and layout, and by subtle changes to answer-sheets. These concerns about context effects have caused at least some high-stakes test-publishers to go to great lengths to standardize these and other aspects of the testing experience. In the first instances of large-scale high-stakes uses of CAT, this strict adherence to standardization carried over to computers, which were also standardized so that the computer hardware and software were virtually identical across test administrations for a given exam. This strict adherence to standardization meant that testing could occur only on specially designated computers. Consequently, their cost was factored into the cost of CAT testing.

If adaptive testing could be routinely conducted on computers used for other purposes (and this could be done without loss of precision or validity), then a primary impediment of CAT testing (i.e., hardware costs) could be substantially reduced. Some test developers are for example considering the administration of low or medium stakes adaptive tests

over the Internet, thus enabling CAT to be administered on a wide variety of general-purpose computer platforms. Because of the important role computer hardware plays in both the economic and psychometric viability of CAT, necessary research on context effects is likely to intensify in the coming years. This includes research on the aspects of the computer hardware that influence test performance, and on characteristics of the exam (such as speededness) and software-interface that might interact with particular hardware characteristics. Progress in these and related areas is likely to further increase the popularity of computerized adaptive testing.

Further Reading

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Weiss, D. J. (n.d.). *CAT Central: A global resource for computerized adaptive testing research and applications.* Retrieved December 12, 2003, from http://www.psych.umn.edu/ psylabs/ CATCentral/.