From the *User's Manual for the MicroCAT Testing System (Second Edition)*, St. Paul: Assessment Systems Corporation, Copyright 1987.

CHAPTER 2

A Brief Introduction To Computerized Testing

2.1 Introduction

A psychological test is defined by the manner in which items are selected for presentation and by the method used to compute a score from an examinee's responses to those items. The way items are chosen is called the item-selection strategy, and the way scores are computed from item responses is called the scoring method.

There are two major kinds of testing strategies: conventional and adaptive. A *conventional test* is constructed by selecting a fixed set of items for administration to a group of individuals. A conventional test is typically scored by counting the number of items that are answered correctly. Such tests are simple to design and score, but they are not particularly efficient because the same set of items is administered to everyone regardless of ability. For any particular individual, many of the items may be much too easy or much too hard, providing little information for pinpointing that person's standing with regard to others of similar ability. Tests of this type are inefficient in their use of administration time and in the number of items needed to provide an accurate estimate of each examinee's ability.

Adaptive tests, on the other hand, are efficient even for a group of individuals differing widely in ability. Adaptive tests are based on a simple concept: more information can be obtained from a test item if the item is matched to the ability level of the examinee. To discriminate among low-ability examinees, relatively easier items should be administered; to discriminate among high-ability examinees, relatively more difficult items should be administered. Practical complications of adaptive testing arise from two sources: 1) an examinee's ability level must be known in advance in order to choose the most appropriate items, and 2) when everyone answers a different set of items, the test cannot be scored by simply counting the number answered correctly.

Another type of testing strategy supported by the system is the *individualized domain-referenced test.* In domain-referenced testing, items are randomly sampled from domains and performance on the items is assumed to be representative of the performance that would be observed if all of the items in the domain were administered. Most domain-referenced tests are constructed by sampling items from the domain, constructing a conventional test using these items, and administering the test to individuals. The

MicroCAT Testing System supports this kind of testing procedure with the Conventional Test Building Program in its Conventional Testing Subsystem. The system also allows you to administer individualized domain-referenced tests. In these tests, subsets of items are randomly selected and administered at the time of testing and each individual responds to a different subset. Because each test is different (subject to any limitations imposed by the size of the item pool), the test is not a standard conventional one. Because the test is not tailored to the ability or some other characteristic of the examinee, it is not really an adaptive test either, as defined above.

Another testing strategy supported by the system is *fixed-branching simulation.* In this type of testing strategy, the items administered are dependent upon one or more earlier responses by the examinee. It is not an adaptive testing strategy, as defined above, because branching is not necessarily dependent upon the correctness or incorrectness of the examinee's response(s). For example, a fixed-branching simulation of managerial effectiveness might present a situation to the examinee and ask the examinee to choose an appropriate action. The next item presented would depend on the action chosen, and the situation presented in it would reflect the effects of the action.

The MicroCAT Testing System supports and facilitates all of these testing strategies. Because adaptive testing is less familiar to many practitioners than the other testing strategies, the remainder of this chapter will describe adaptive testing in more detail. A bibliography at the end of the chapter suggests additional sources for information.

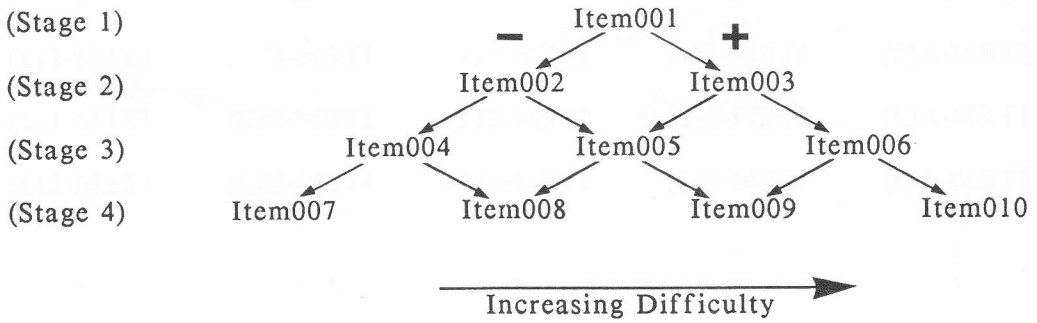
2.2 Intuitively Branched Adaptive Strategies

The solution to the first problem of adaptive testing (selecting items without knowing an examinee's ability level) has been approached intuitively from several directions. The simplest solution is to create a hierarchy of subtests. Scores from one test are used to estimate the examinee's ability level, and then subsequent tests are selected on the basis of the score on the first one. One of the earliest of these hierarchical strategies is the *two-stage test* (Angoff & Huddleston, 1958; Betz & Weiss, 1973, 1974; Lord, 1980; Weiss, 1974) in which all examinees first respond to a common routing test. The score on that test is then used to assign each examinee to a second-stage measurement test. Responses to both tests are used to arrive at a final score. A problem with the two-stage strategy is that errors in measurement on the first-stage test result in misrouting to an inappropriate measurement test at the second stage.

The problem of misrouting in the first stage led to the development of multistage tests in which a common routing test leads to two or more

second-stage routing tests which, in turn, lead to multiple third-stage measurement tests. The addition of extra routing tests reaches its logical limit when only one item remains in each subtest. The most popular example of this limiting case is called the *pyramidal test* (Bayroff, Thomas, & Anderson, 1960; Larkin & Weiss, 1974; Lord, 1970; Weiss, 1974). In a pyramidal test, everyone starts with the same item and then branches either to an easier item after each incorrect response or to a more difficult item after each correct response. A diagram of a pyramidal test is shown in Figure 2-1.

Figure 2-1. Diagram of a Pyramidal Test



As the number of items at each stage of testing increases, the pyramidal structure begins to make very inefficient use of its items. While a 15-item two-stage test might require 45 items, a 15-item pyramidal test requires 120. This problem led to the development of re-entrant strategies in which a subtest containing a given set of items could be partially administered, exited for a subtest at another level, and then returned to if necessary. The best example of this type of strategy is the *stradaptive test* (Vale & Weiss, 1975a, 1975b, 1978; Weiss, 1973) in which several (for example, nine) subtests (or strata) are defined, each containing items at a specified difficulty level. In this strategy, testing proceeds by administering an item in one stratum and then branching to a more difficult stratum if the item is answered correctly or to a less difficult stratum if it is answered incorrectly. Whenever testing branches to a stratum, the next previously unadministered item in that stratum is administered to the examinee. A diagram of the structure for a five-stratum (A-E) stradaptive test is shown in Figure 2-2.

Several such mechanical branching mechanisms were evaluated over the years, but no strategy had a clear psychometric advantage over all of the others. However, solutions to the second practical problem in adaptive testing (test scoring) have produced some superior strategies. While a few of the intuitively branched strategies described above could be scored in simple and meaningful ways, many could not. The general solution to both the item-selection and the test-scoring problems lay in item response theory.

Figure 2-2. Diagram of a Five-Stratum Stradaptive Test With a Pool of 5n Items

STRATUM A	STRATUM B	STRATUM C	STRATUM D	STRATUM E
ITEM-A(1)	ITEM-B(1)	ITEM-C(1)	ITEM-D(1)	ITEM-E(1)
ITEM-A(2)	ITEM-B(2)	ITEM-C(2)	ITEM-D(2)	ITEM-E(2)
ITEM-A(3)	ITEM-B(3)	ITEM-C(3)	ITEM-D(3)	ITEM-E(3)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
ITEM-A(n)	ITEM-B(n)	ITEM-C(n)	ITEM-D(n)	ITEM-E(n)

2.3 Statistically Branched Adaptive Strategies

2.3.1 Item Response Theory Models

Item response theory (IRT) is a statistical theory consisting of a family of models that express the probability of observing a particular response to an item as a function of certain characteristics of the item and of the ability level of the examinee (Hambleton & Swaminathan, 1985). IRT models have several forms depending on the format of the item response options and the simplifying assumptions made regarding the process underlying the response.

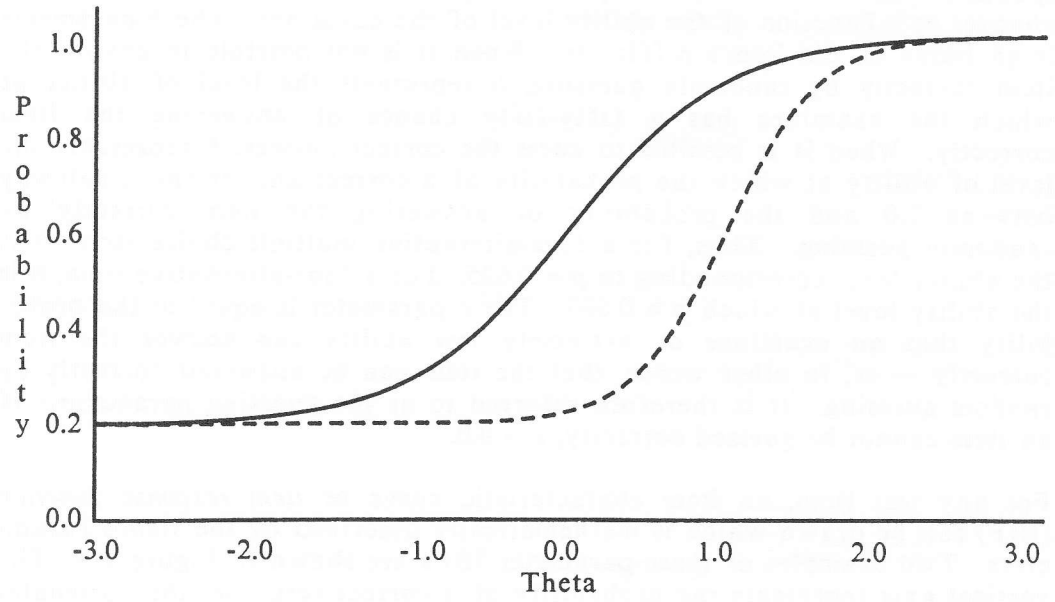
The model that has been used most widely in computerized adaptive testing (CAT) is the three-parameter logistic model. This model is applicable to multiple-choice questions scored in a dichotomous (for example, correct-

incorrect) manner. It describes the probability of observing a correct response to an item as a function of the examinee's ability level (called *theta*) and three item parameters (a, b, and c). The a parameter represents the item's capability of discriminating between levels of ability. It corresponds to the rate at which the probability of a correct response to an item changes as a function of the ability level of the examinee. The b parameter is an index of the item's difficulty. When it is not possible to answer the item correctly by randomly guessing, b represents the level of ability at which the examinee has a fifty-fifty chance of answering the item correctly. When it is possible to guess the correct answer, b represents the level of ability at which the probability of a correct answer (p) is halfway between 1.0 and the probability of answering the item correctly by randomly guessing. Thus, for a five-alternative multiple-choice item, b is the ability level corresponding to $p = 0.625$. For a four-alternative item, b is the ability level at which $p = 0.600$. The c parameter is equal to the probability that an examinee of extremely low ability can answer the item correctly — or, in other words, that the item can be answered correctly by random guessing. It is therefore referred to as the guessing parameter. If an item cannot be guessed correctly, $c = 0.0$.

For any test item, an item characteristic curve or *item response function* (IRF) can be drawn which is mathematically described by the item's parameters. Two examples of three-parameter IRFs are shown in Figure 2-3. The vertical axis represents the probability of a correct response; the horizontal axis represents theta, the examinee's ability level. Thus, an item's IRF illustrates the probability of getting the item correct as a function of an examinee's ability level. The shape of the IRF depends upon the item's parameters.

The IRF drawn with the solid line in Figure 2-3 represents an item with $a = 1.0$, $b = 0.0$, and $c = 0.2$. The slope of the curve is related to the item's a (discrimination) parameter. Higher values of a would make the IRF steeper. At theta levels where the IRF is steeper, the item's capability of discriminating among examinees with ability levels near that level of theta is increased. The location of the curve along the horizontal axis is a function of the b (difficulty) parameter. Higher values of b shift the curve to the right; lower values shift the curve to the left. The lower left asymptote of the IRF corresponds to the item's c (guessing) parameter. Higher values of c raise the asymptote, indicating an increased probability of correct answers for examinees with very low ability levels. In Figure 2-3, the IRF drawn with a dashed line represents an item with $a = 2.0$, $b = 1.0$, and $c = 0.2$. The lower asymptote remains at 0.2. Note, however, that the higher value of b has shifted the midpoint of the curve to a theta level of 1.0. Because a is also larger, the IRF for this item is steeper than the IRF for the item with $a = 1.0$.

Figure 2-3. Two Three-Parameter Item Response Functions



The three-parameter model can be considered a general model for dichotomously scored items. It allows items to differ in their discriminating powers, in their difficulties, and in how easily they can be answered correctly by guessing. However, this general form requires much computation to estimate the three parameters as well as substantial computation to obtain scores. If items can be assumed not to vary on all of these characteristics, computational savings can be obtained by setting some of the parameters to constant values.

The guessing parameter, *c*, causes the most computational difficulty and can be set to 0.0 if the items cannot be answered correctly by guessing. Recall-type items that do not provide an opportunity for guessing can be used with this reduced model. The model with *c* assumed to be 0.0 is called the two-parameter model.

The discrimination parameter, *a*, can also be set to a constant if it is reasonable to assume that all items are equally good at discriminating high abilities from low abilities. The one-parameter model is usually referred to

incorrect) manner. It describes the probability of observing a correct response to an item as a function of the examinee's ability level (called *theta*) and three item parameters (a, b, and c). The a parameter represents the item's capability of discriminating between levels of ability. It corresponds to the rate at which the probability of a correct response to an item changes as a function of the ability level of the examinee. The b parameter is an index of the item's difficulty. When it is not possible to answer the item correctly by randomly guessing, b represents the level of ability at which the examinee has a fifty-fifty chance of answering the item correctly. When it is possible to guess the correct answer, b represents the level of ability at which the probability of a correct answer (p) is halfway between 1.0 and the probability of answering the item correctly by randomly guessing. Thus, for a five-alternative multiple-choice item, b is the ability level corresponding to $p = 0.625$. For a four-alternative item, b is the ability level at which $p = 0.600$. The c parameter is equal to the probability that an examinee of extremely low ability can answer the item correctly — or, in other words, that the item can be answered correctly by random guessing. It is therefore referred to as the guessing parameter. If an item cannot be guessed correctly, $c = 0.0$.

For any test item, an item characteristic curve or *item response function* (IRF) can be drawn which is mathematically described by the item's parameters. Two examples of three-parameter IRFs are shown in Figure 2-3. The vertical axis represents the probability of a correct response; the horizontal axis represents theta, the examinee's ability level. Thus, an item's IRF illustrates the probability of getting the item correct as a function of an examinee's ability level. The shape of the IRF depends upon the item's parameters.

The IRF drawn with the solid line in Figure 2-3 represents an item with $a = 1.0$, $b = 0.0$, and $c = 0.2$. The slope of the curve is related to the item's a (discrimination) parameter. Higher values of a would make the IRF steeper. At theta levels where the IRF is steeper, the item's capability of discriminating among examinees with ability levels near that level of theta is increased. The location of the curve along the horizontal axis is a function of the b (difficulty) parameter. Higher values of b shift the curve to the right; lower values shift the curve to the left. The lower left asymptote of the IRF corresponds to the item's c (guessing) parameter. Higher values of c raise the asymptote, indicating an increased probability of correct answers for examinees with very low ability levels. In Figure 2-3, the IRF drawn with a dashed line represents an item with $a = 2.0$, $b = 1.0$, and $c = 0.2$. The lower asymptote remains at 0.2. Note, however, that the higher value of b has shifted the midpoint of the curve to a theta level of 1.0. Because a is also larger, the IRF for this item is steeper than the IRF for the item with $a = 1.0$.

as the Rasch model (Rasch 1980; Wright & Stone, 1979), named after its developer.

While the three-parameter logistic IRT model is appropriate only for dichotomously scored items, other IRT models are available for use with other response formats (for example, Bock, 1972; Samejima, 1969). Polychotomous models, for instance, are appropriate where multiple response categories are scored on each item. One example of this situation is a multiple-choice item in which the incorrect alternatives are weighted as a function of how "incorrect" they are. Another example is an interest-test item with ordered Like, Indifferent, and Dislike responses. Still another is a performance-rating scale on which performance is rated in ordered categories.

IRT models generally assume one of two shapes for the IRF. Most models assume that the response probabilities follow a logistic ogive (a specific shape of the IRF). Early in the development of IRT, several models were based on a normal ogive rather than a logistic ogive. The normal ogive model arose from the widespread use of the normal curve for statistical models in psychology. The shape of the IRFs is nearly the same for both models and it is difficult to say which fits reality better. The logistic ogive is more attractive because of its mathematical simplicity, and it has replaced the normal model in most practical implementations.

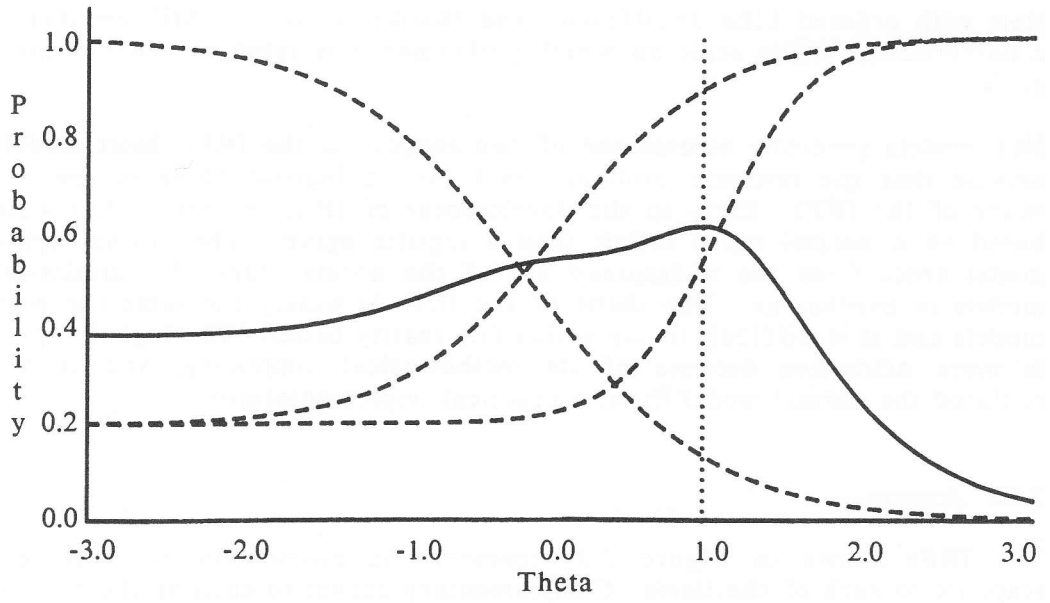
2.3.2 Scoring

The IRFs shown in Figure 2-3 represent the probability of a correct response to each of the items. Complementary curves to each of these exist, representing the probability of an incorrect response. The local-independence assumption of IRT (that performance on a particular item is independent of success or failure on other items) allows the curves corresponding to the correct and incorrect responses in an examinee's entire test response pattern to be multiplied together to yield a *likelihood function.* The likelihood function indicates the probability of observing the entire vector of obtained item responses at each level of ability. From this likelihood function, an estimate of the examinee's ability can be obtained. Conceptually, this can be done by assuming that the best estimate of an examinee's ability is the level of ability that would most likely produce the vector of responses observed. This is determined by locating the maximum value of the likelihood function and identifying the ability level (theta) associated with that maximum. This score is called the *maximum likelihood estimate of ability.*

Figure 2-4 shows three IRFs as dashed lines for three items, two answered correctly and one answered incorrectly. The solid curve shows the product of these curves (that is, the likelihood function). The point at which the

likelihood function reaches its peak (the vertical dotted line) corresponds to a theta value of approximately 0.9. This value of theta is the maximum likelihood estimate of the examinee's ability.

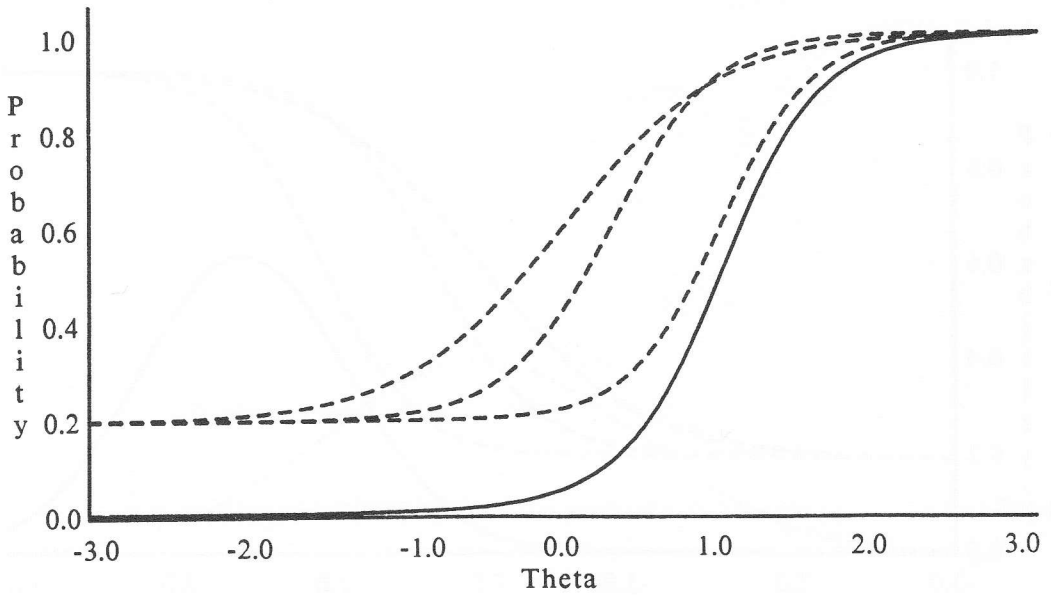
Figure 2-4. Three IRFs and a Likelihood Function With Two Items Answered Correctly and One Item Answered Incorrectly



One problem with maximum likelihood estimation is that it occasionally produces estimates of ability at positive or negative infinity on the theta scale. An example of this is shown in Figure 2-5. Again, three items were presented. This time, however, the examinee answered all three correctly. The likelihood function (the solid line) has no obvious peak since the curve continues to rise imperceptibly as theta increases beyond 3.0. The maximum likelihood estimate is, therefore, positive infinity.

Several methods of bounding maximum likelihood estimates have been used, many of them practical but arbitrary. An alternative to the maximum likelihood method is the *Bayesian modal method*. A Bayesian modal estimate is conceptually very similar to the maximum likelihood estimate; in fact, it is simply an extension of it. It differs in that a Bayesian prior likelihood function is included when the IRFs are multiplied together. This eliminates

Figure 2-5. Three IRFs and a Likelihood Function With All Items Answered Correctly

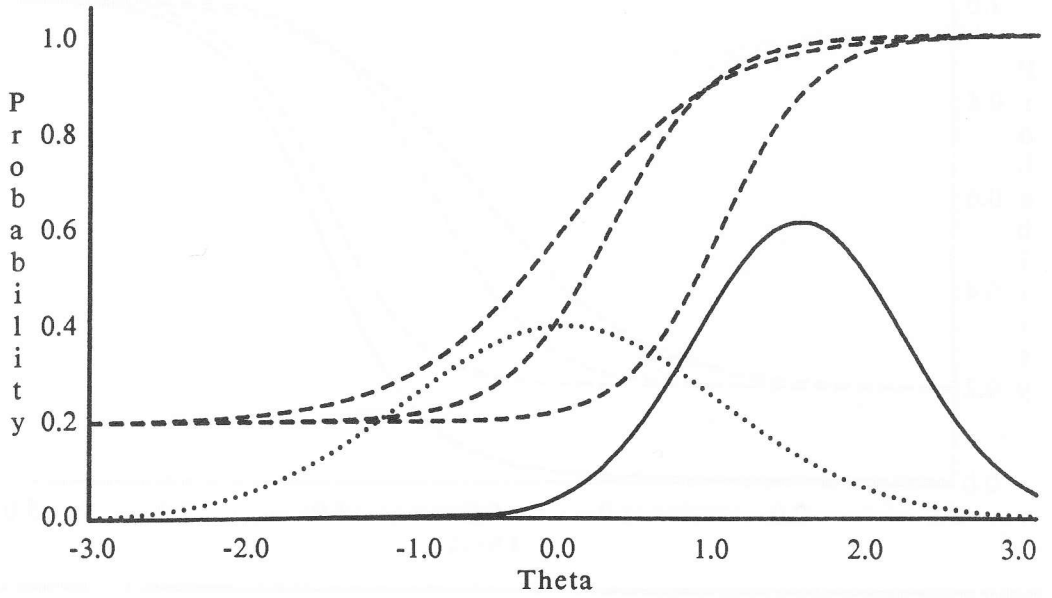


the infinite estimates. Figure 2-6 shows the effect on the IRFs used in Figure 2-5 of assuming that the distribution of theta (that is, the prior likelihood function) is standard normal, as shown by the dotted curve. A finite peak of the modified likelihood function (the solid line) now exists at a theta value of approximately 1.5.

2.3.3 Item Selection

Maximum likelihood and modal Bayesian ability estimation can be used with any adaptive testing item-selection strategy when the items are calibrated according to an IRT model. These scoring methods have suggested some useful and flexible item-selection strategies. Because a more peaked likelihood function yields a more accurate estimate of ability, it makes good sense to explicitly select test items that will sharpen the peak of the likelihood function.

Figure 2-6. Three IRFs, a Normal Bayesian Prior Distribution for Theta, and a Bayesian Posterior Likelihood Function With All Items Answered Correctly



Item information is a statistical concept closely related to the slope of the IRF and inversely related to the standard error of the ability estimate that would result if the item were administered and scored. Item information curves are transformations of IRFs and, like IRFs, they are a function of theta and the parameters of the item. The *maximum information item-selection strategy* selects items based on item information curves. First, an estimate of ability is obtained. The information value for each item is then evaluated at that level of theta. The item with the highest value of information at that theta level is chosen as the best item to administer. In a maximum information adaptive test, a sequential process is specified in which an item is administered, an ability estimate is calculated, the item providing the most information at that estimate is selected, and the process is repeated. The sequential process may continue until a fixed number of items has been administered or until some other criterion for termination (such as a specified value of the standard error of the ability estimate) has been satisfied.

The *Bayesian item-selection strategy* is similar to the maximum information strategy, except that it selects items on the basis of minimizing the Bayesian posterior variance of the ability estimate rather than maximizing values of item information. Because of the relationships between information and the Bayesian posterior variance, however, these item-selection strategies will frequently select the same items. See Weiss (1982) for a discussion of the relationships between these item-selection methods, as well as examples of the use of adaptive testing for different testing problems.

2.4 Summary

The MicroCAT Testing System supports and facilitates conventional testing via computer, computerized adaptive testing, individualized domain-referenced testing, and fixed-branching simulation testing. The MicroCAT system is the only commercially available, general-purpose testing system for the development, administration, and scoring of computerized adaptive tests.

Computerized adaptive testing (CAT) is a method of constructing maximally efficient tests by tailoring the items included in a test to the examinee's ability. A CAT strategy consists of a method for selecting items and a method for scoring the responses. Many CAT strategies have been developed and each may be useful in a particular context. This chapter provided some basic information about CAT which, while not replacing further study of and experience with CAT, should give some understanding of the procedures involved.

2.5 References and Additional Reading

Angoff, W. H. & Huddleston, E. M. (1958). *The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test* (Statistical Report SR-58-21). Princeton, NJ: Educational Testing Service.

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

Betz, N. E. & Weiss, D. J. (1973, October). *An empirical study of computer-administered two-stage ability testing* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. E. & Weiss, D. J. (1974, October). *Simulation studies of two-stage ability testing* (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Clark, C. L. (Ed.). (1976, March). *Proceedings of the First Conference on Computerized Adaptive Testing.* Washington, DC: US Civil Service Commission, Personnel Research and Development Center.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boaston: Kluwer-Nijhoff.

Larkin, K. C. & Weiss, D. J. (1974, July). *An empirical investigation of computer-administered pyramidal ability testing* (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

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

McBride, J. R. (1979, November). *Adaptive mental testing: The state of the art* (Tech. Report 423). Alexandria: US Army Research Institute for the Behavioral and Social Sciences.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago: University of Chicago Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2).

Vale, C. D. & Weiss, D. J. (1975a, October). *A study of computer-administered stradaptive ability testing* (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Vale, C. D. & Weiss, D. J. (1975b, December). *A simulation study of stradaptive ability testing* (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Vale, C. D. & Weiss, D. J. (1978). The stratified adaptive ability test as a tool for personnel selection and placement. *TIMS Studies in the Management Sciences*, *8*, 135-151.

Warm, T. A. (1978, October). *A primer of item response theory* (Tech. Report CG-941278). Oklahoma City: US Coast Guard Institute, Department of Transportation.

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

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

Weiss, D. J. (Ed.). (1978, July). *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (Ed.). (1980, September). *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

Weiss, D. J., & Vale, C. D. (in press). Adaptive testing. *International Review of Applied Psychology.*

Weiss, D. J., & Vale, C. D. (in press). Computerized adaptive testing for measuring abilities and other psychological variables. In J. N. Butcher (Ed.), *The Practitioner's Guide to Computer-Based Psychological Testing.* New York: Basic Books.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: Mesa Press.