Computerized Adaptive Testing with a Military Population

Steven Gorman
Headquarters, U.S. Marine Corps

Adaptive testing is a computer-assisted interactive process which facilitates the rapid, accurate measurement of the ability of the testee. The process begins with an examinee, seated at a terminal, being presented a question and responding to that question. After each response, an estimate of the examinee's ability is updated. The computer program then selects a question which is either more difficult or easier. One way to select the next question is to choose, out of a pool of available questions, one that will minimize the standard error of estimate of the examinee's ability when it is administered. This tailoring process, with a minimum number of test items, maximizes the information obtained about an individual's ability level.

With the mathematically elegant Bayesian algorithm developed by Owen (1975) in conjunction with a cathode ray tube and computer interface, adaptive testing is now not only possible, but also necessary for large-scale testing situations, such as military accession testing. This mandate is founded upon the potential benefits of adaptive testing:

1. Greater test precision at all ability levels, especially at the tails of the distribution;
2. Improved test security;
3. Decreased misclassification;
4. Reduction of examinee anxiety or boredom;
5. Reduction in test length;
6. Enhanced discrimination of testee abilities;
7. Enhanced applicant motivation as a result of immediate feedback on test responses;
8. Standardized test administration;
9. Improved data quality through elimination of human requirements for calculation and data recording; and
10. Interface with classification, assignment, and job information systems.

Adaptive testing is based upon latent trait theory--the theory which analyzes examinee performance at the item level. Accordingly, it is extremely important that items be carefully screened for use in adaptive testing and that the item parameters be satisfactorily estimated. The assumptions and procedures for "norming" items in adaptive testing require greater information than was the case with norming tests under classical test theory. These requirements and procedures have been stated previously (Urry, 1976, Jensema, 1976a). A discussion of these requirements and the effects of their

violations is in order; this paper will subsequently present a discussion of the item parameterization, the adaptive testing research design, and the metric base for each procedure.

A Brief Overview of Theory

The only appropriate model for representing multiple-choice items is the three-parameter latent trait model (Birnbaum, 1968; Urry, 1971). This model can be represented in either a normal ogive or the more popular logistic function:

$$P'(\theta) = c_i + (1-c_i)P, \qquad [1]$$

where

$$P = \frac{1}{\sqrt{2\pi}} \int_{a_i(\theta-b_i)}^{\infty} e^{\frac{-x^2}{2}}\, dx \cong \frac{1}{1 + exp[-Da_i(\theta - b_i)]} \qquad [2]$$

a_i = the item discriminating power;

b_i = the inflection point of the item characteristic curve, or the item difficulty;

c_i = the lower asymptote of the regression of item response on the latent trait, also referred to as the guessing parameter;

D = the constant 1.7; and

θ = the latent trait continuum of ability, which ranges from $-\infty$ to $+\infty$, but usually is restricted to the range -3 to +3.

Figure 1 depicts an item characteristic curve where a_i=2.0, b_i=0.0, and c_i=.18. The curve is based on a plot of $P'(\theta)$, the probability of successfully answering a test question, given ability level θ, when guessing is effective. The value c_i=.18 may occur in a multiple-choice test item that has five response alternatives. Note that the c_i value is less than .20. This may be attributable to the attractiveness of the wrong alternatives at greater than chance level. The value b_i=0.0 occurs at the probability location P=.5(1+.18)=.59. The value a_i is related to the slope at the point b_i, which is the inflection point of the curve.

Figure 2 depicts $Q(\theta)$, the curve displaying the probability of having ability θ, given that an examinee responded incorrectly to an item. As can be seen in Figure 2, the slope of $Q(\theta)$ is steeper than $P'(\theta)$, shown in Figure 1.

Figure 1
Item Characteristic Curve (P')

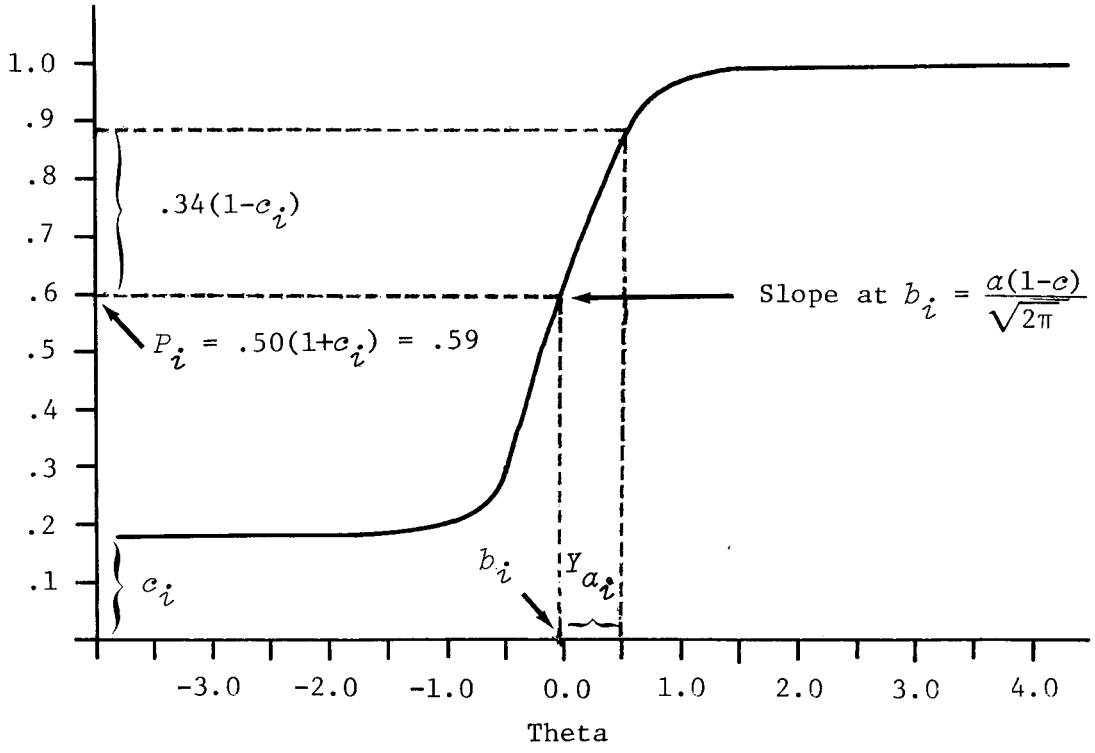
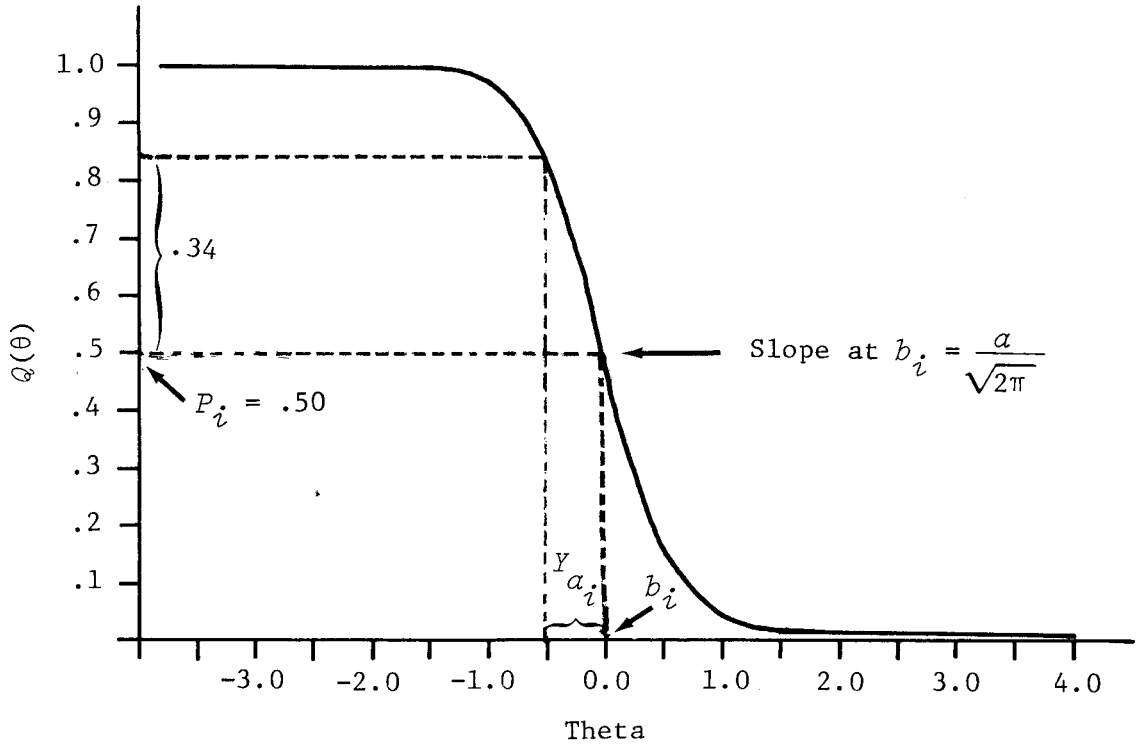


Figure 2
Curve for Wrong Response (Q)



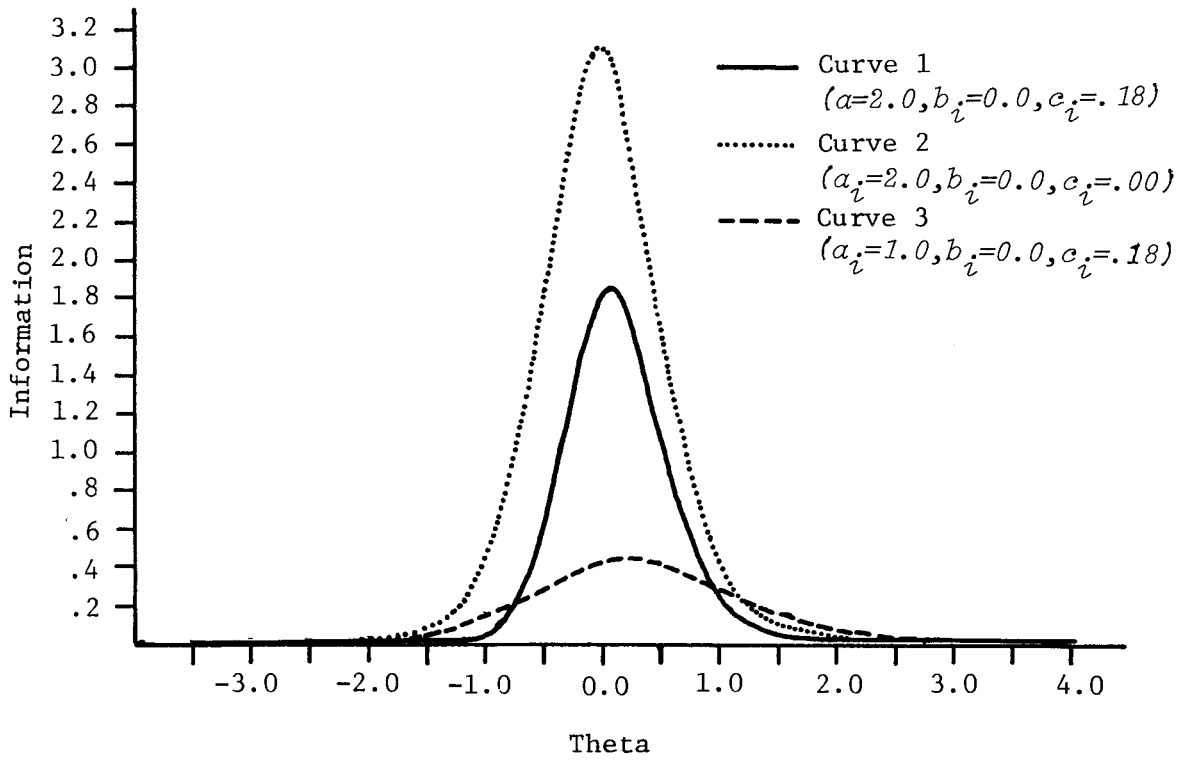
As will be shown more clearly later, an incorrect response to an item decreases the standard error of estimate more rapidly than a correct response.

Birnbaum (1968) developed the concept of information in a test item. The simple formula is

$$I(\theta) = \sum \left(\frac{[\text{slope at } \theta]^2}{\text{variance}} \right) \qquad [3]$$

Figure 3 displays the information available in three test items. Curve 1 is based on the item with a_i=2.00, b_i=0.0, and c_i=.18 which is displayed in Figure 1. Curve 2 has the same a_i and b_i parameters, but c_i=.00. Curve 3 has the parameters a_i=1.0, b_i=0.0, and c_i=.18. Notice that high values of a_i and low values of c_i increase the amount of information available in an item. Also of interest is the fact that when c_i is non-zero, the information curve is skewed (reflecting the fact that guessing is effective) and thus has a greater effect on the lower end of the curve.

Figure 3
Information Curves



Research Plan

Selection of Test Items

Test items which best measure Word Knowledge (WK) and Arithmetic Reasoning (AR) abilities will be selected. A minimum of 180 WK and 180 AR items will be required. The criteria for pre-screening these items include:

1. Highest discriminatory power;
2. Difficulty values over the full ability range; and
3. Five alternative responses per item, if available.

The actual procedure involves "guesstimating" the item coefficient of guessing (c_i)=.20, and computing either the conventional point-biserial or biserial correlation between item-total test score and the proportion passing. Utilizing the charts provided in the literature (Urry, 1974, 1977), approximations of the item parameters are obtained. Jensema (1976) provides a computer program to generate approximations for item parameters using conventional item statistics.

Development of Parallel Tests

The test items will be divided into parallel test forms containing 60 test questions. The test forms will be generated by distributing the approximations for the item difficulties (b_i) widely and evenly throughout the ability range. This procedure will ensure a test information function which is reasonably flat over the entire ability range; this is important, since it provides a base for the manifestation of latent ability.

Failure to include a minimum of 60 test items or failure to measure all levels of ability evenly will provide an inadequate base to measure latent ability, resulting in inferior item parameter estimates. Monte carlo simulation (Gugel et al., 1976) has shown that 60 items is the minimum required for estimating reliable item parameters.

Administration of Test Booklets

Test booklets will be administered to a minimum of 2000 incoming Marines at the recruit depots. The examinees will be instructed to answer all 60 questions and to use as much time as necessary. The population at the recruit depots covers a full ability range of the target population (those personnel whose age makes them available for military service), thereby assuring that the parameters are properly located. Monte carlo simulation (Gugel et al., 1976) demonstrated that utilization of 2000 or more examinees will yield sufficiently reliable item paramter estimates.

If the test is treated as a speeded test rather than as a power test, two actions would ensue: (1) items at the end of the test would not be completed by the slower examinees; and (2) responses to items would be rushed, creating higher item discriminating powers (a_i), but lower item difficulties (b_i).

Estimating parameters with a population which is not representative of the target population will result in estimates of a_i and b_i which are algebraic transformations of the target population parameters. Using anchor items with both populations will enable determination of the differences in ability distribution. The algebraic transformation can be determined from the means and standard deviations of the two populations.

Estimation of Item Parameters

The item responses will be transferred from the answer sheets to computer input media. The item vectors will be processed via computer, utilizing the Urry (1976) item parameter estimation technique. This program cycles through two successive approximations of estimating the distribution of manifest ability: (1) corrected raw scores where the item being parameterized is omitted, and (2) using Bayesian modal ability estimates.

The output provides estimates of the parameters a_i, b_i, and c_i, using a smallest chi-square fit. After both cycles the approximations are then adjusted for the information of the group of items to increase their efficiency.

Adaptive Testing Item Bank Selection

The item bank requirements for efficient and accurate tailoring (Jensema, 1976a; Urry, 1974, 1977) are:

1. Item discriminatory powers (a_i) should be as high as possible, and at least .80.
2. Item difficulties (b_i) should be widely and evenly distributed.
3. Item coefficients of guessing (c_i) should be as low as possible, with .30 as a maximum.
4. Item bank should consist of a sufficient number of items.

Jensema (1976) has shown that when values of a_i become large and when value of c_i become small, greater tailoring efficiency occurs. This conclusion is consistent with the finding that greater information is available when an item has high a_i values and low c_i values. Where there are gaps in the distribution, failure to satisfy the requirement for a rectangular difficulty distribution will force the Bayesian algorithm to select an item which will yield less than optimal information about the examinee's ability.

Evaluation of Adaptive Testing Item Bank

An evaluation of the effectiveness of the adaptive testing item bank will be conducted through a simulation run of approximately 500 computer-generated "examinees." This will be accomplished by a monte carlo simulation of a normal (0,1) population for θ, using the estimated item parameters and the Bayesian ability estimation program for multiple ability banks. This procedure will allow an evaluation of the performance of the item banks in adaptive testing and will serve as a guide for assessing the future performance of testing live examinees.

Adaptive Test Administration

A computer-administered adaptive ability test will be administered to a sample of recruits at the Marine Corps Recruit Depot. This test will measure WK and AR abilities, using the Bayesian ability estimation program for multiple ability banks. The termination rule will be a pre-selected reliability, using the formula:

$$\sigma_\varepsilon = \sqrt{1-\rho^2_{\theta\hat{\theta}}} \qquad [4]$$

Using a standard error of estimate as a termination rule guarantees equiprecision of ability estimation for all examinees. Use of a maximum number of items as the termination rule will guarantee lower test reliability for upper ability examinees. The reason for this is that the Bayesian ability estimation process assumes a normal prior distribution of ability for all examinees. The effect of this prior distribution is shown in Figures 4 through 7.

Figure 4
$P'(\theta)$ (a_i=2.09, b_i=-.12, c_i=.17)
and Posterior Distribution

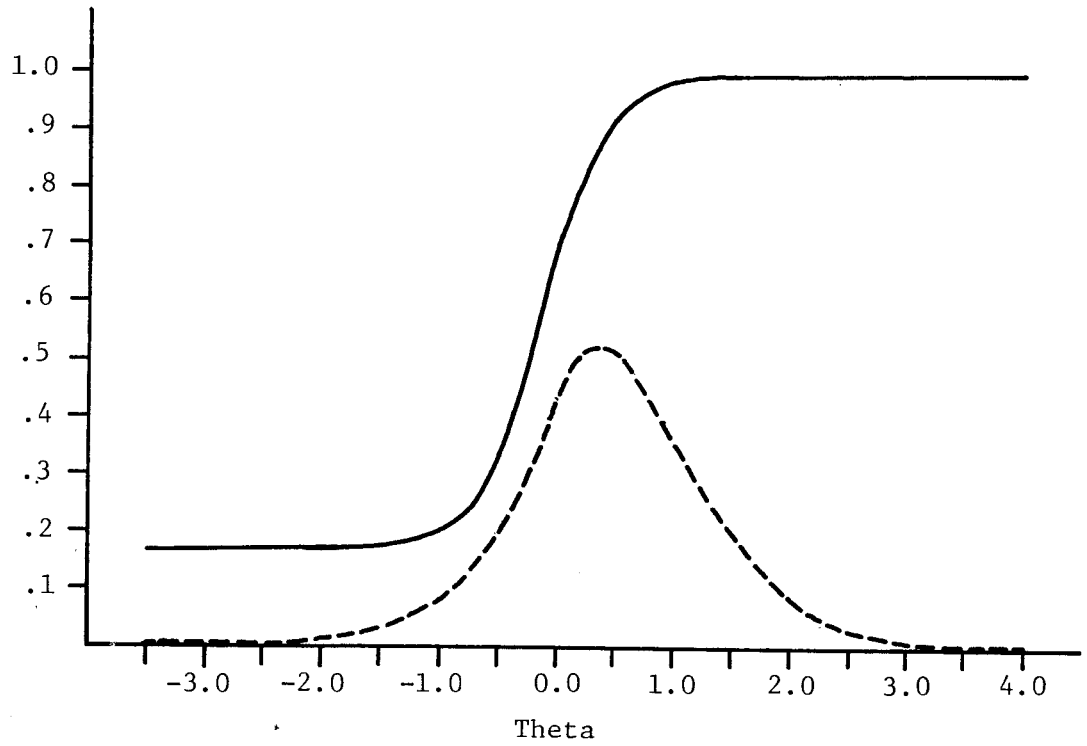
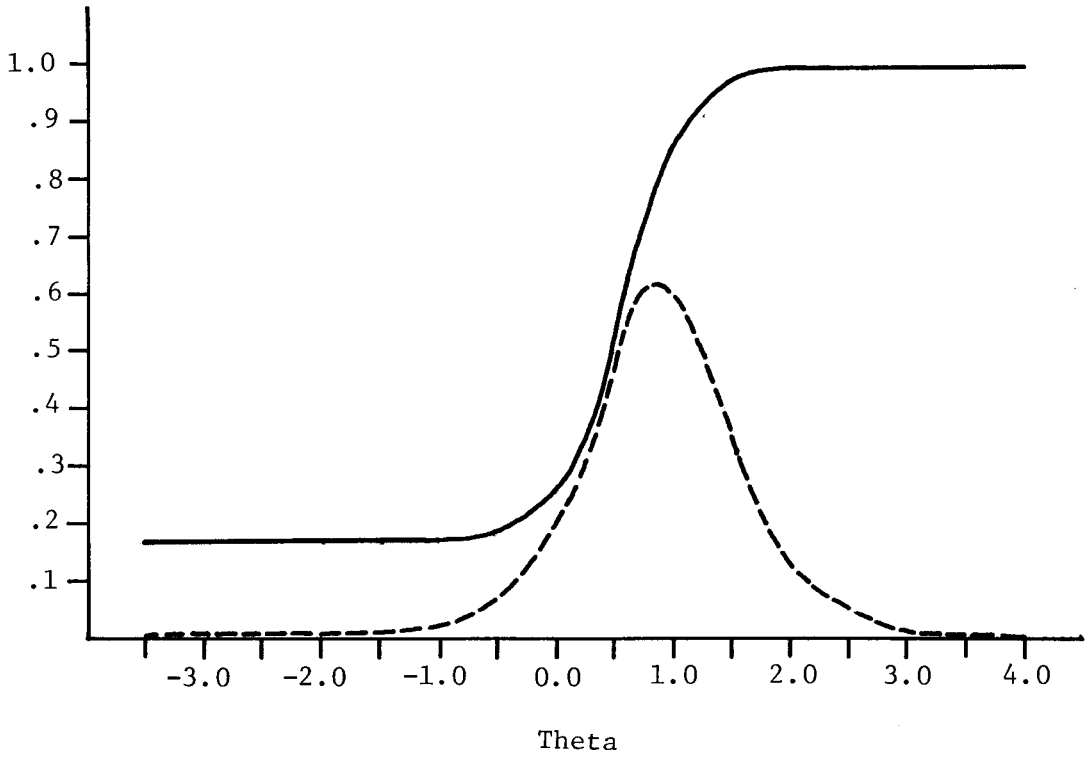


Figure 4 shows the posterior distribution after an examinee correctly responded to an item with a_i=2.09, b_i=-.12, and c_i=.17. At this point the ability estimate was .469, with standard error of .8568. Figure 5 shows the

posterior distribution after the examinee had correctly responded to an item with a_i=2.25, b_i=.59, and c_i=.17. The ability estimate was then updated to θ=.929, with standard error = .7527. Figures 6 and 7 illustrate the greater capacity of an incorrect response to decrease the standard error of estimate. In Figure 6 the examinee responded incorrectly to an item with a_i=1.94, b_i=.78, and c_i=.13. The ability estimate was then .367 with standard error of .5549.

Figure 5
$P'(\theta)$(a_i=2.25, b_i=.59, c_i=.17)
and Posterior Distribution



Had the examinee responded correctly, as in Figure 7, the standard error would have decreased only to .6440. Since lower ability applicants will have more incorrect responses, their standard errors of estimate will decrease more rapidly. Those examinees who respond correctly to a greater proportion of items will require a longer test in order to maintain equiprecision of ability estimation.

Figure 6
$Q(\theta)$(a_i=1.94, b_i=.78, c_i=.13)
and Posterior Distribution

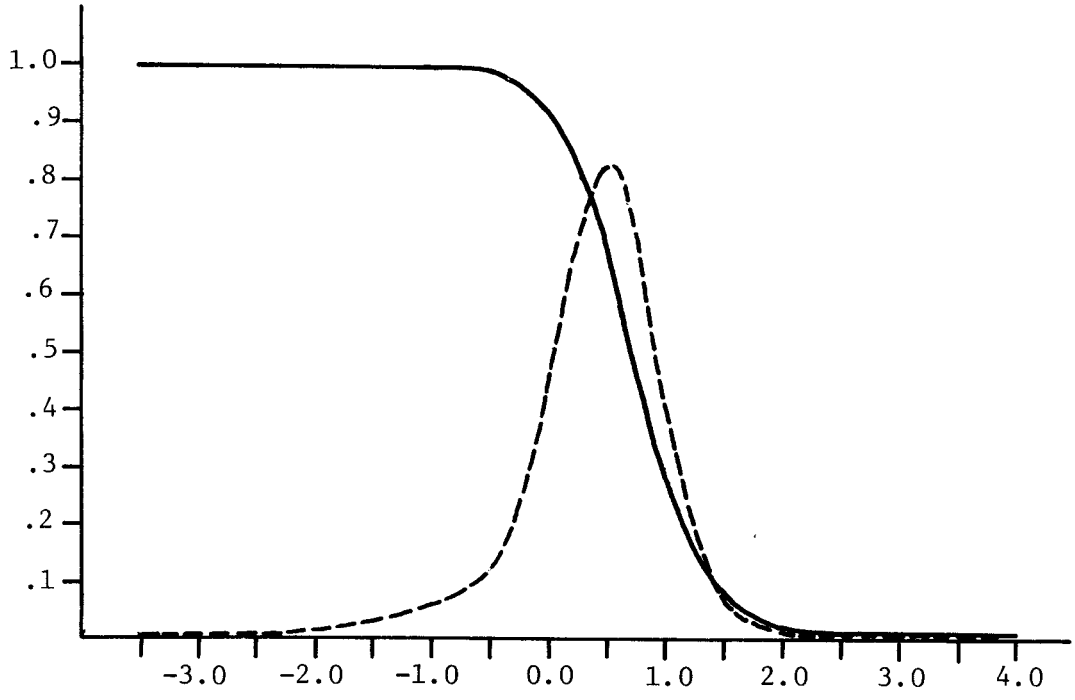
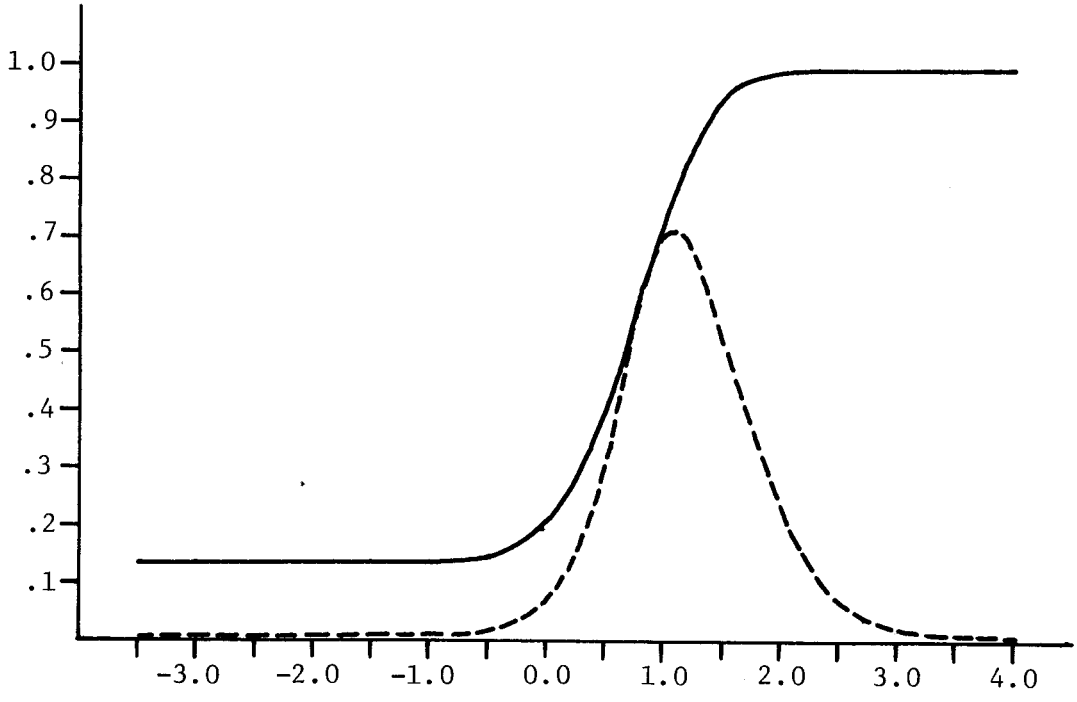


Figure 7
$P'(\theta)$(a_i=1.94, b_i=.78, c_i=.13)
and Posterior Distribution



Summary

This paper has demonstrated some of the errors which can be committed in instituting Bayesian adaptive testing. With an awareness and sensitivity to these and other potential problems, psychometricians can gain numerous benefits by using computerized adaptive testing.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Gugel, J. F., Schmidt, F. J., & Urry, V. W. Effectiveness of the ancillary estimation procedure. Proceedings of the first conference on computerized adaptive testing (PS-75-6). Washington, DC: Personnel Research and Development Center, U. S. Civil Service Commission, 1976.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: Personnel Research and Development Center, U. S. Civil Service Commission, 1976. (a)

Jensema, C. J. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 1976, 36, 705-715. (b)

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

Urry, V. W. A monte carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970). Dissertation Abstracts International, 1971, 31, 6319B. (University Microfilms No. 71-9475).

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? Proceedings of the first conference on computerized adaptive testing (PS-75-6). Washington DC: Personnel Research and Development Center, U. S. Civil Service Commission, 1976.

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