AD-A272 832

Extracting Information From Wrong Answers in Computerized Adaptive Testing

DTIC
ELECTE
NOV 18 1993
A
S

J. Bradford Sympson

93-28290

93 11 17 004

Extracting Information From Wrong Answers
in Computerized Adaptive Testing

J. Bradford Sympson

DTIC QUALITY INSPECTED 5

Reviewed by
Daniel O. Segall

Accesion For		
NTIS CRA&I		☑
DTIC TAB		☐
U announced		☐
Justification		
By		
Distribution /		
Availability Codes		
Dist	Avail and / or Special	
A-1		

Approved and released by
W. A. Sands
Director, Personnel Systems Department

Approved for public release;
distribution is unlimited.

Navy Personnel Research and Development Center
San Diego, California 92152-7250

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE October 1993	3. REPORT TYPE AND DATE COVERED Final—October 1988-September 1991

4. TITLE AND SUBTITLE
Extracting Information From Wrong Answers in Computerized Adaptive Testing

5. FUNDING NUMBERS
Program Element: 0601153N
Work Unit: R4204

6. AUTHOR(S)
J. Bradford Sympson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
San Diego, California 92152-7250

8. PERFORMING ORGANIZATION REPORT NUMBER
NPRDC-TN-94-1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Assistant Secretary of Defense (FM&P)
The Pentagon
Washington, DC 20301-3210

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Functional Area: Personnel
Product Line:　Computerized Testing
Effort:　Computerized Adaptive Testing

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (Maximum 200 words)

A brief review of the history of polychotomous (i.e., multi-category) item response models is provided. After describing a new polychotomous item response model (Model 8), examples of the Operating Characteristic Functions obtained when Model 8 is applied to real test data are given. In general, inspection of "goodness-of-fit" plots indicates that Model 8 provides superior data fit and higher item information functions than the well-known 3-parameter logistic (dichotomous) item response model. A simulation of computerized adaptive testing (CAT) that used the actual item responses of applicants for military enlistment shows that Model 8 would be superior to the 3-parameter logistic model in a CAT environment. In this investigation, Model 8 increased test reliability by an amount that is equivalent to a 25% increase in test length.

14. SUBJECT TERMS
Computerized testing, selection, classification, training, testing, item response theory, psychometric models, Computerized Adaptive Testing-Armed Services Vocational Aptitude Battery, CAT-ASVAB, item scoring, polychotomous scoring, polytomous scoring

15. NUMBER OF PAGES
33

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UNLIMITED

Foreword

This technical note provides a brief history of the development of polychotomous (i.e., multi-category) item-response models, describes a new model developed by the author, and reports the results of a data analysis in which both dichotomous and polychotomous item-response models are applied to test data obtained from applicants for military enlistment.

W. A. SANDS
Director, Personnel Systems Department

Summary

Problem

Conventional methods for scoring aptitude and achievement tests that are used in selecting, classifying, and training military personnel discard useful information about an examinee's ability/skill level. Information is lost whenever the original responses to test questions are classified only as "right" or "wrong." Additional information can be obtained by considering the difficulty level of the questions answered correctly and by taking into account which particular wrong answers were selected.

Objective

The objective of this effort was to develop new procedures for scoring aptitude and achievement tests that will increase the reliability and validity of those tests.

Approach

Enhanced scoring procedures previously developed in the field of psychometrics (psychological and educational testing) were reviewed. Emphasis was placed on methods developed under Item Response Theory (IRT). A new polychotomous (i.e., multi-category) IRT model was developed. An existing dichotomous (right/wrong) IRT model and the new polychotomous IRT model were both applied to test data obtained from applicants for military enlistment. The data were analyzed in a way that simulated the effects of scoring computerized adaptive tests (CAT) with these two IRT models.

Results

Inspection of graphic plots of the item-response data showed that the dichotomous IRT model could not accurately track the observed proportions of correct responses for some items. Simulation of CAT showed that the new polychotomous scoring procedure increased test reliability by an amount that was equivalent to a 25% increase in test length.

Conclusions

The new polychotomous IRT model developed in this research provides a superior foundation for scoring computerized tests. The new model should provide even larger reliability increases when applied to paper-and-pencil tests that have previously been scored with conventional methods such as number/proportion-correct. Users of the new scoring method can elect either to keep personnel tests at their current length and increase score reliability, or to reduce test length in order to save testing time while maintaining score reliabilities at current levels.

Recommendation

Organizations that administer aptitude and/or achievement tests for purposes of personnel selection, classification, or training should consider whether the new IRT model and scoring method described in this research can be usefully applied to their tests.

Contents

List of Figures

Introduction

Procedures for the development and use of personnel tests are now being revised to incorporate the concepts and methods derived from Item Response Theory (IRT) (Lord, 1980). IRT provides a logical basis for improving conventional testing methods and also provides a foundation for the development of new testing methods (e.g., Computerized Adaptive Testing [CAT]).

The principal characteristic that distinguishes IRT from traditional testing theory is the use of stochastic models that can be used to compute the probability that examinees of a given ability level will answer a test question in a specified manner. At this time, there are two stochastic models that dominate practical applications of IRT, the 1-parameter logistic model (Rasch, 1961) and the 3-parameter logistic model (Birnbaum, 1968).

The 1-parameter logistic model has been adopted in a variety of psychometric applications. The popularity of this model derives from its conceptual simplicity and the relative ease with which the model can be fit to item-response data. Unfortunately, in the last 10 years evidence has accumulated that indicates this model is not adequate for calibrating multiple-choice items (e.g., see Hambleton & Murray, 1983). The 3-parameter logistic model is now being used in an increasing range of applications as more test theorists and practitioners become familiar with its characteristics and as better methods for fitting the model to item response data become available.

Both of the IRT models described above have two drawbacks. First, both models are dichotomous models; they classify responses to multiple-choice test questions as either correct or incorrect. They do not distinguish among the different incorrect responses a person might select. Information about a person's level of ability that could be extracted by taking into account which particular incorrect responses have been selected is lost when these models are used.

Second, these models assume that the probability of a correct response to a test question is a strictly increasing function of ability. There is empirical evidence that this assumption is false for a portion of the test questions in many content domains. For example, in 1980, Sympson examined data from 125 Verbal questions and 145 Quantitative Reasoning questions that had appeared in operational forms of the *Scholastic Aptitude Test* (SAT). For most of the questions, the proportion of correct responses tended to increase monotonically as ability increased. However, for about one question in six, the proportion of correct responses tended to decline as ability increased from very low levels to moderate levels, and then to increase as ability went from moderate to high levels. This finding can be attributed to the presence in some of the questions of one or more "plausible" incorrect responses that were more attractive to examinees at middle ability levels than to examinees at low ability levels.

Darrell Bock[1] also found evidence of non-monotone correct-response probabilities when he analyzed data obtained from eight different tests of the *Armed Services Vocational Aptitude Battery* (Department of Defense, 1984). Both Michael Levine and Fumiko Samejima[2] found similar evidence in their analyses of other aptitude test items.

---

[1]Personal communication, 1983.

[2]Personal communications, 1985-1986.

If a non-trivial portion of the test questions in a given content domain have non-monotone regressions of the type described above, then it is clear that currently-used IRT models cannot provide an adequate basis for item calibration in that domain. Models that can accommodate non-monotonicity, when it does occur, are needed.

Samejima (1979) published the first report describing such models. Samejima described generalizations of previously developed IRT models for both graded and nominal item responses. Samejima's models for graded responses will not be discussed here, since their range of potential application is more limited than that of nominal models. Samejima's nominal model is a generalization of a model originally proposed by Bock (1972). Bock's model was appropriate for situations in which guessing was discouraged and examinees could be depended upon to omit a question if they did not feel they knew the correct answer. These highly specialized requirements served to eliminate Bock's model from consideration in most practical applications.

Samejima's nominal model can be derived from Bock's model under the assumption that people who would omit a question, if allowed to do so without penalty, will guess at random among the available item responses whenever they are not allowed to omit. Both common sense and available empirical evidence suggest that this random guessing assumption is not viable (e.g., see Strang, 1977, and references cited therein). However, the fact that Samejima's model can be derived from a restrictive model and an implausible assumption does not imply that the model will necessarily be inadequate as a tool for use in psychometric applications. This is an empirical question that can only be answered by accumulation of relevant data.

Sympson (1981) proposed the first nominal model in which response-category Operating Characteristic Functions (OCFs) could approach different limiting values as ability decreased. Whereas Samejima eliminated the "omit" response category and assumed that the remaining response categories would have equal limiting probabilities as ability decreased, Sympson proposed to retain the "omit" response category and to estimate the limiting probability associated with each possible response. However, Sympson did not offer any empirical evidence to refute Samejima's "equal-lower-limit" assumption and did not offer any evidence that the necessary additional model parameters could, in practice, be estimated.

Thissen and Steinberg (1983) described a polychotomous model that is formally equivalent to the model proposed by Sympson (1981). They found that the parameters of their model (hence, equivalently, Sympson's 1981 model) could, in practice, be estimated and that this model provided significantly better fit to their data than did Samejima's (1979) model.

It may be useful to point out some logical relationships among the IRT models that have been described so far. Bock's (1972) nominal model is a polychotomous extension of the 2-parameter (dichotomous) logistic model. Sympson's (1981) nominal model (and, hence, Thissen and Steinberg's model) is a polychotomous extension of the 3-parameter (dichotomous) logistic model. Samejima's (1979) nominal model is a polychotomous version of the 3-parameter (dichotomous) logistic model with an auxiliary assumption that all OCFs approach a limit equal to the reciprocal of the number of response alternatives as ability decreases without bound.

2

All of the IRT models discussed so far have one key feature in common. They are based on mathematical expressions in which the relationship between underlying ability and the probability of observing a particular response derives from an exponential function in which the argument (i.e., the exponent of the irrational number e) is a linear function of ability. The first departure from this assumption of "linear logits" was proposed by Sympson (1983). On the basis of an analysis of the logical implications of using various mathematical functions for the logit, Sympson concluded that a more appropriate polychotomous model would incorporate a non-decreasing cubic-polynomial logit for the correct-response category, a concave-downward quadratic-polynomial logit for each of the incorrect alternatives, and a linear logit for the "omit" response.

Sympson (1983) used large-sample SAT data to compare the fit of his "Model 6" to the fit of previously proposed polychotomous models and found that the new model provided significantly better fit. In particular, Sympson's results showed that neither Samejima's model nor Sympson's earlier (1981) model was adequate for the items studied. Sympson's results also showed that it was possible to fit a relatively complex IRT model (one with 23 parameters per item) without encountering intractable numerical difficulties.

Sympson (1983) also examined the fit of two other polychotomous models that had not previously been proposed, but that were less complex than the full model he derived on the basis of logical considerations. Sympson's results suggested that the complexity of the full model was needed in order to obtain an adequate level of fit to the data analyzed.

Later analyses by Sympson revealed that Model 6, while superior to previous polychotomous IRT Models, failed to fit a few of the items examined. This led to the development of a somewhat more flexible model referred to as Model 8 (Sympson, 1986a, 1986b, 1986c).

Examples of Polychotomous Operating Characteristic Functions

Sympson's Model 8 gives the probability of selecting response category j ($j = 1, 2, \ldots, m$) as a function of π, where π is a continuous, real-valued measure of ability that falls on the closed interval [0,1]. Note that $j = 1$ corresponds to response option "A," $j = 2$ corresponds to response option "B," etc. This mapping was selected for notational convenience and does not imply an ordering of the response categories as is required by graded item-response models.

Under Model 8, the probability of choosing response option j is given by the expression

$$P_j[\pi] = \frac{e^{f_j}}{\sum_{k=1}^{m} (e^{f_k})} , \qquad (1)$$

where

$$f_j = p_{5j-4} + p_{5j-3}\pi + p_{5j-2}\pi^2 + p_{5j-1}\pi^3 + p_{5j}\pi^4 . \qquad (2)$$

Thus, in this model, each "logit" is a quartic polynomial in π.

In Equation 2, lower-case italic *p* indicates a model parameter. Associated with each parameter is a subscript of the form 5*j-i*, where *i* ranges from zero to 4. Each parameter subscript indicates both the response category with which the parameter is associated and the particular parameter under consideration.

In general, there are five parameters per response category. However, the parameters for any one category can all be set equal to zero during item calibration. Thus, for a 5-choice item, if omitting is treated as a sixth response category, 25 parameters are estimated and 5 parameters are fixed. In practice, the parameters for the correct-response category should be fixed at zero.

Model 8 has been fit to 86 multiple-choice vocabulary items that were previously calibrated with the 3-parameter logistic (3PL) model. Since each item had five response options and omitting was treated as a separate category, 25 parameters were estimated for each item. Item response data from 2,607 applicants for military service that had been used in obtaining the 3PL OCFs were used to obtain the Model 8 OCFs. To obtain a rank-ordering of the examinees with respect to ability, an optimal scaling procedure was used (Sympson, 1984). Then, after forming 50 ability groups of approximately equal size (*N* = 52 or 53), the non-linear regression program *BMDP3R* (Dixon, 1981, chapters 14.1 and 14.3) was used to obtain maximum likelihood estimates of the Model 8 parameters for each item. (Jennrich & Moore, 1975, outline procedures for obtaining maximum likelihood parameter estimates from weighted-least-squares fitting algorithms.) For each item, all parameters were estimated simultaneously.

Figures 1 through 5 show fitted OCFs for Item 50, an item that proved to be a good indicator of ability near the middle of the military applicant population. The OCF for Category 6 (omitting) is not displayed here because omitting was discouraged during the original data-collection and occurred only infrequently. The abscissae in these figures are expressed in terms of percentiles. The transformation from [0,1] to [0,100] as an ability metric is straightforward under Model 8.
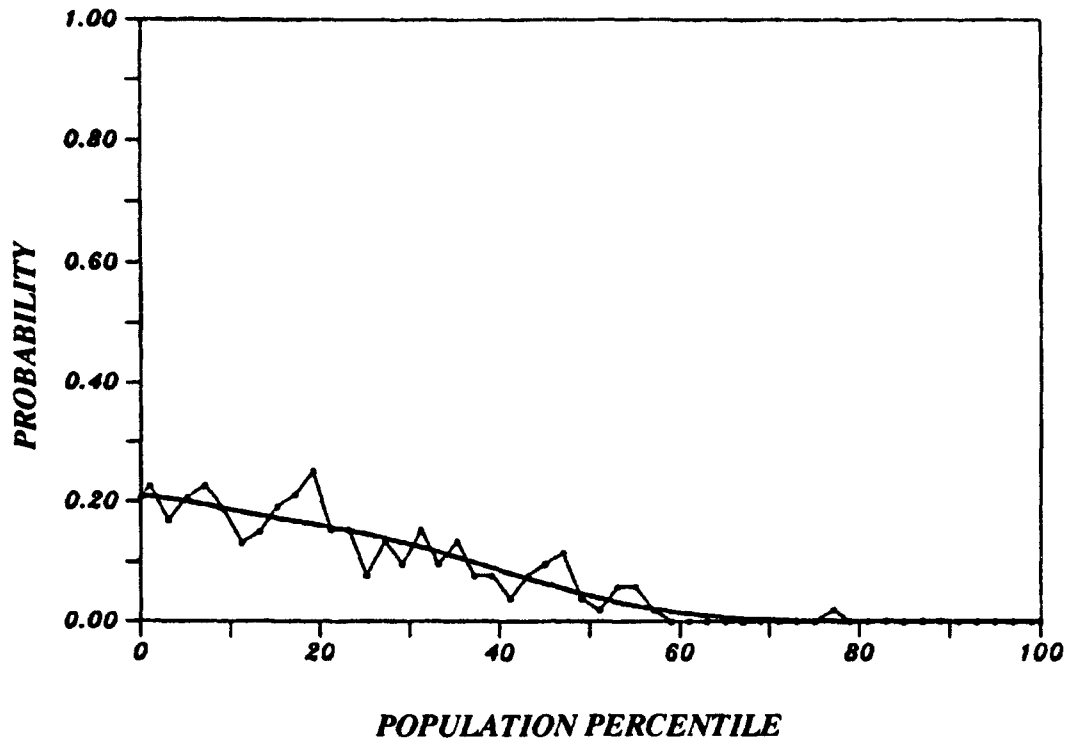
Figure 1. OCF for Item 50, Category 1, under Model 8.
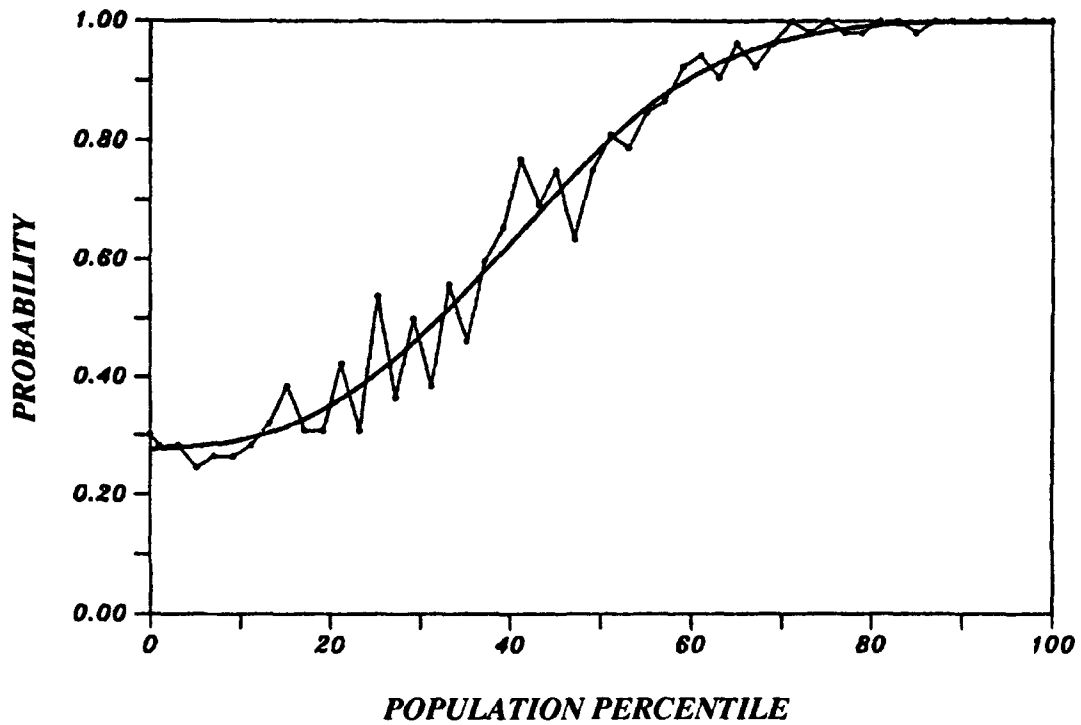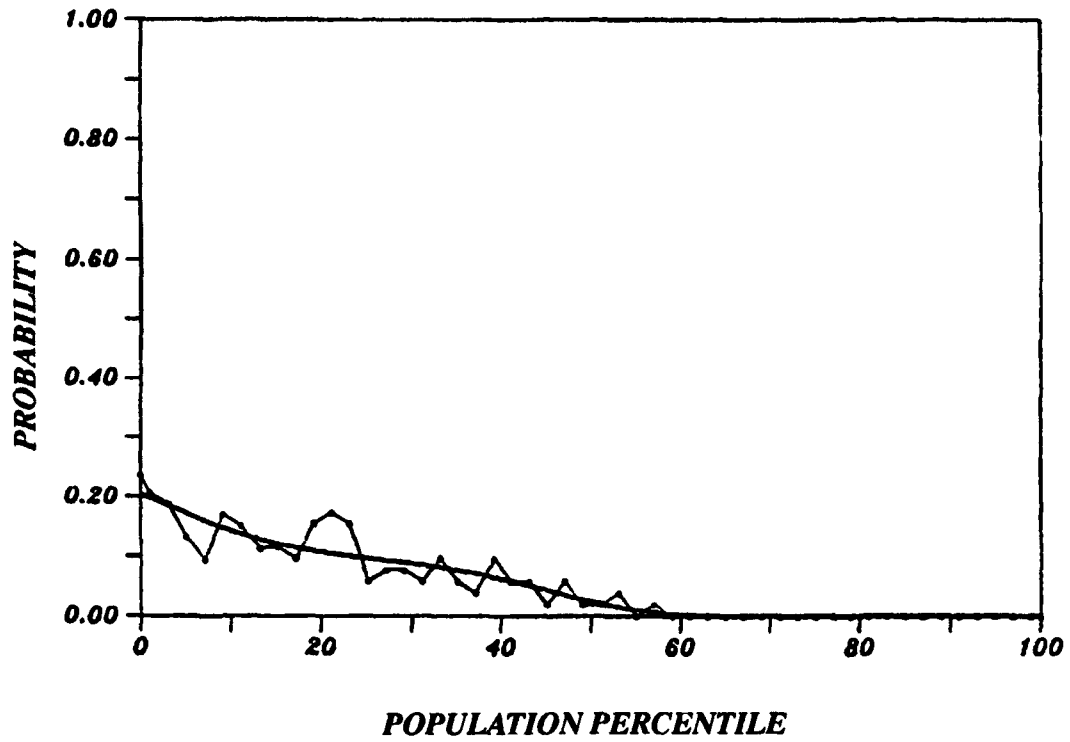

Figure 2. OCF for Item 50, Category 2, under Model 8.

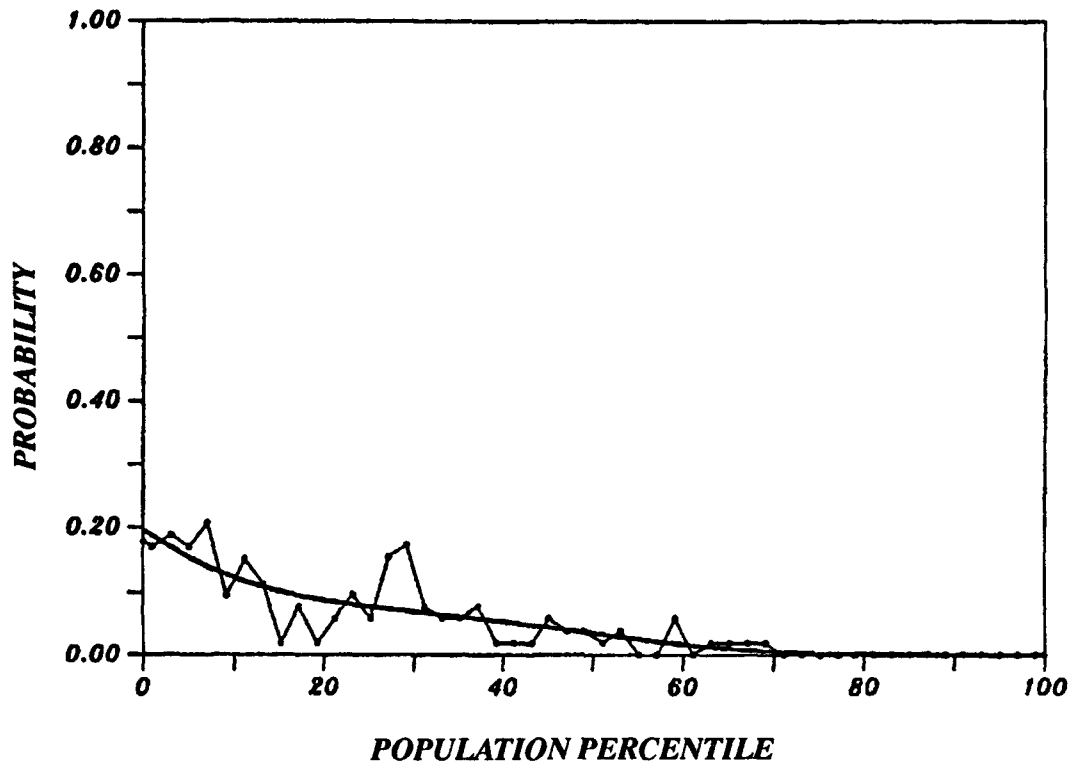**Figure 3. OCF for Item 50, Category 3, under Model 8.**

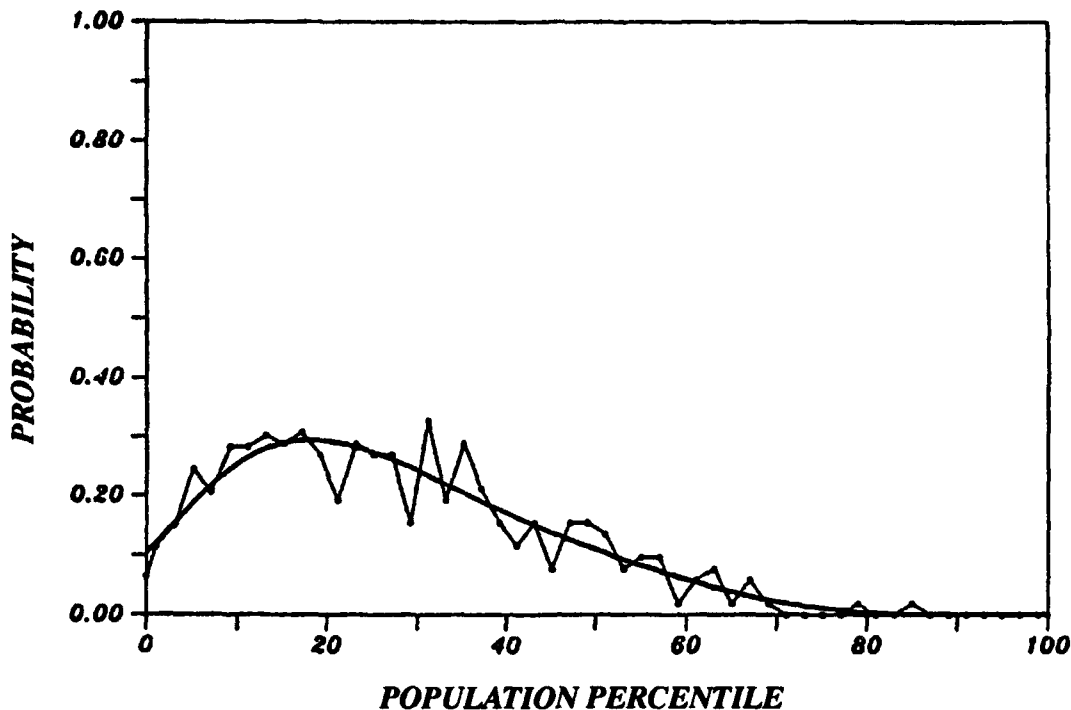**Figure 4. OCF for Item 50, Category 4, under Model 8.**

6

Figure 5. OCF for Item 50, Category 5, under Model 8.

Figure 2 shows the OCF for the correct-response category (option "B") for this item. This OCF indicates that the probability of selecting the correct answer to Item 50 is higher than would be expected by chance (i.e., higher than .20), even at the lowest levels of the military applicant population. This finding is complemented by the fact that the probability of selecting option "E" (Figure 5) is below chance-level at the lowest percentiles.

If π is equal to the population percentile divided by 100, then the density function for π is uniform on [0,1]. If π is transformed to the ability metric θ such that $\theta(\pi)$ has the density function $h(\theta)$, then the *item information function* for item i, with respect to θ, is given by

$$I_i[\theta] = [h(\theta)]^2 I_i[\pi(\theta)]$$

$$= [h(\theta)]^2 \sum_{j=1}^{m} \left\{ P_{ij}[\pi(\theta)] \left[\frac{\partial \ln P_{ij}[\pi(\theta)]}{\partial \pi(\theta)} \right]^2 \right\} . \qquad (3)$$

The information function for an item will be relatively high at ability levels where the item is a good indicator of ability and will be relatively low at ability levels where the item is a poor indicator of ability.

Figure 6 shows two information functions for Item 50 under the assumption that θ is distributed normally with $\mu = 0.0$ and $\sigma = 1.0$. The darker curve in Figure 6 is the item information function under polychotomous scoring of this item. The lighter curve is the item information function under

7

dichotomous scoring of the item. The lighter curve was obtained by collapsing all wrong-answer categories into a single category in Equation 3.
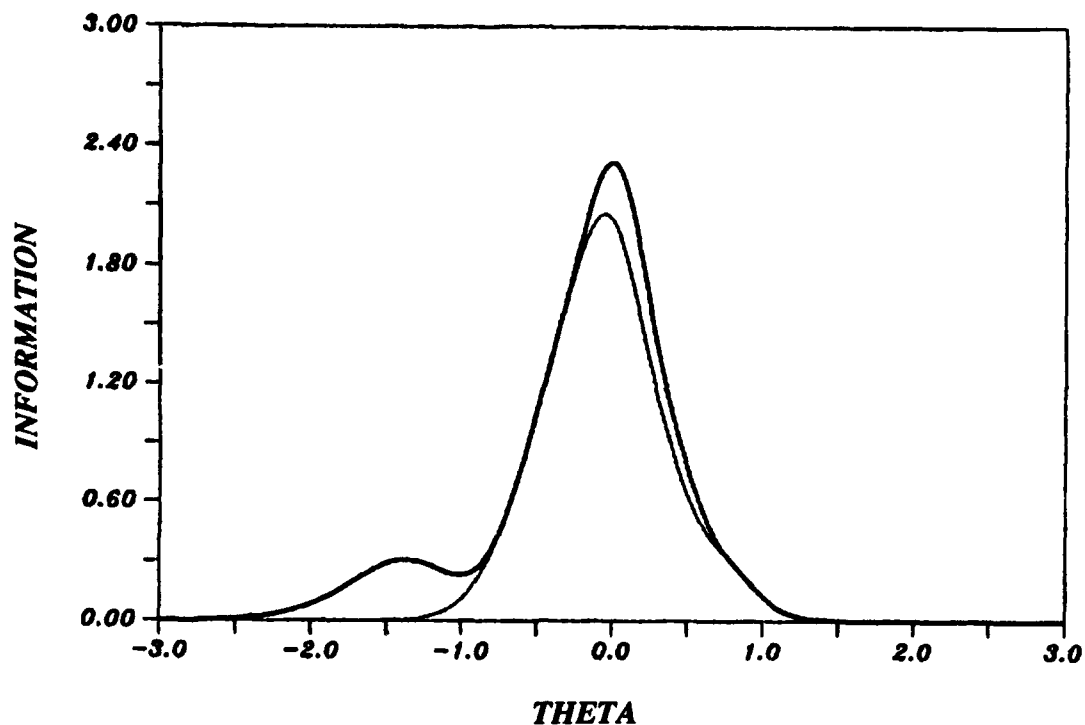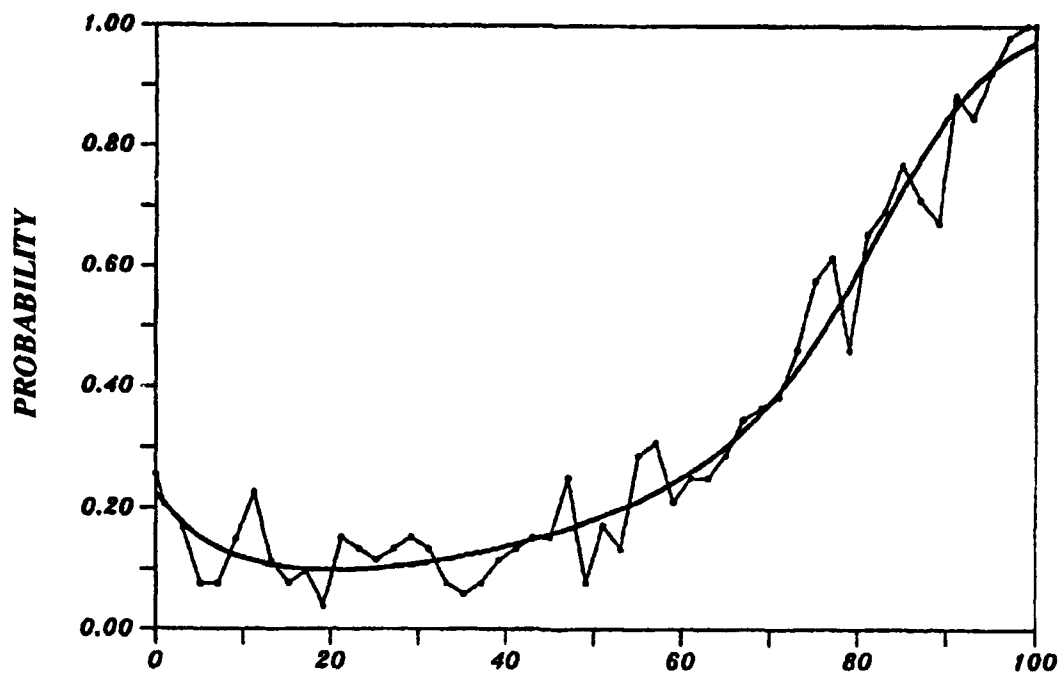

Figure 6. Item information functions for Item 50 under polychotomous and dichotomous scoring.

Figures 7 through 11 show the OCFs for Item 4, a somewhat more difficult item. For this item, Category 1 (Figure 7) is never chosen as much as 20% of the time. However, Category 3 (Figure 9) is chosen with greater than chance frequency over most of the range of this population. The increasing OCF observed for Category 3 below the 15*th* percentile accounts for the non-monotonicity observed in the OCF of the correct-answer category (Figure 8). Figure 12 shows polychotomous and dichotomous item information functions for Item 4 under the assumption that θ is distributed normally with $\mu = 0.0$ and $\sigma = 1.0$.

8

POPULATION PERCENTILE
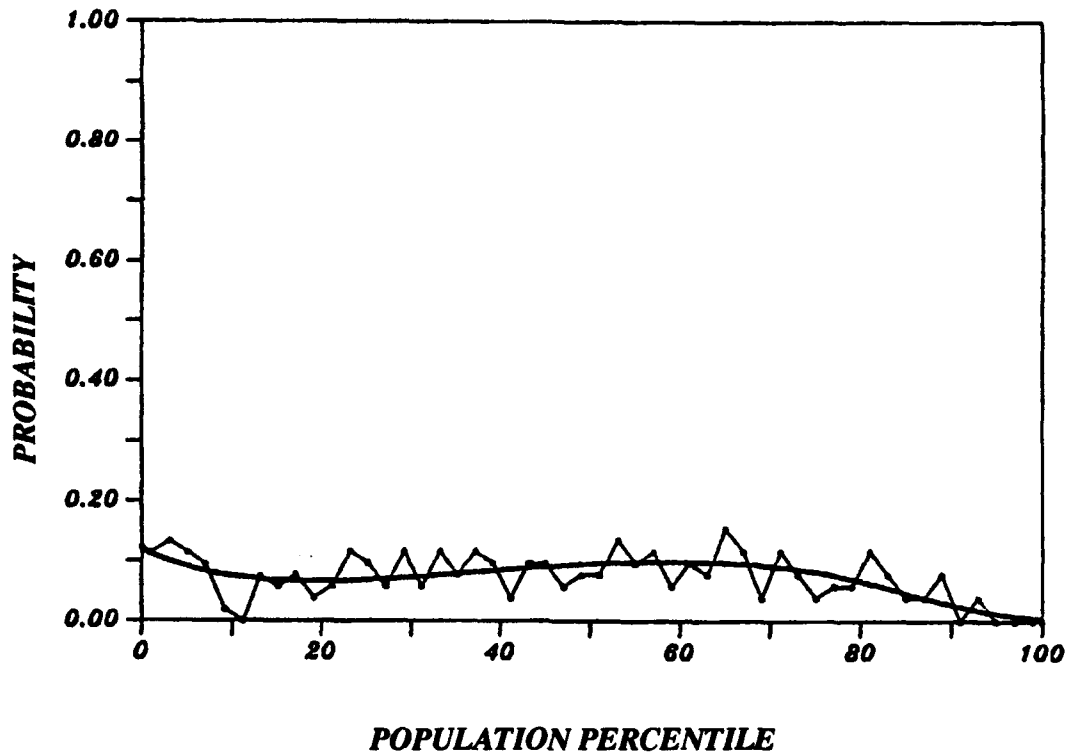
Figure 7. OCF for Item 4, Category 1, under Model 8.

Figure 8. OCF for Item 4, Category 2, under Model 8.

9

POPULATION PERCENTILE

Figure 9. OCF for Item 4, Category 3, under Model 8.

POPULATION PERCENTILE

Figure 10. OCF for Item 4, Category 4, under Model 8.

10

Figure 11. OCF for Item 4, Category 5, under Model 8.

Figure 12. Item information functions for Item 4 under polychotomous and dichotomous scoring.

11

Figures 13 through 17 show the OCFs for Item 6, an easy item. Figure 18 shows the polychotomous and dichotomous item information functions (with respect to normally distributed θ) for this item. When Figures 6, 12, and 18 are compared, *it can be seen that the amount of* additional information available from polychotomous scoring varies from item to item.

Figure 13. OCF for Item 6, Category 1, under Model 8.

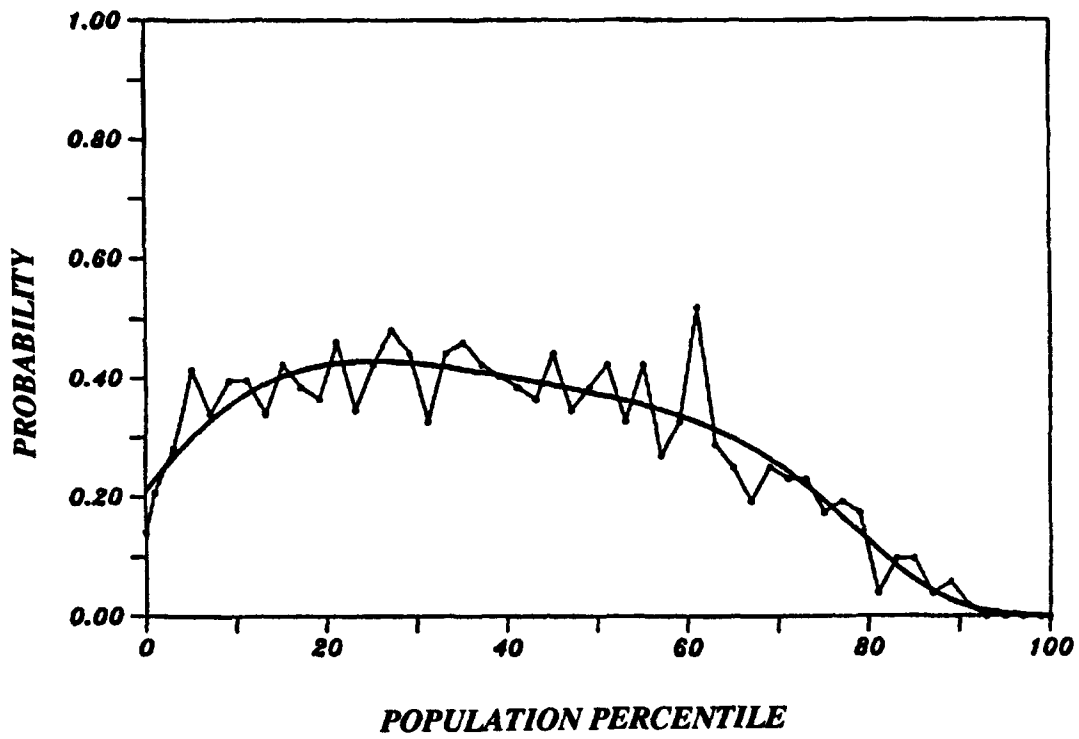**Figure 14. OCF for Item 6, Category 2, under Model 8.**

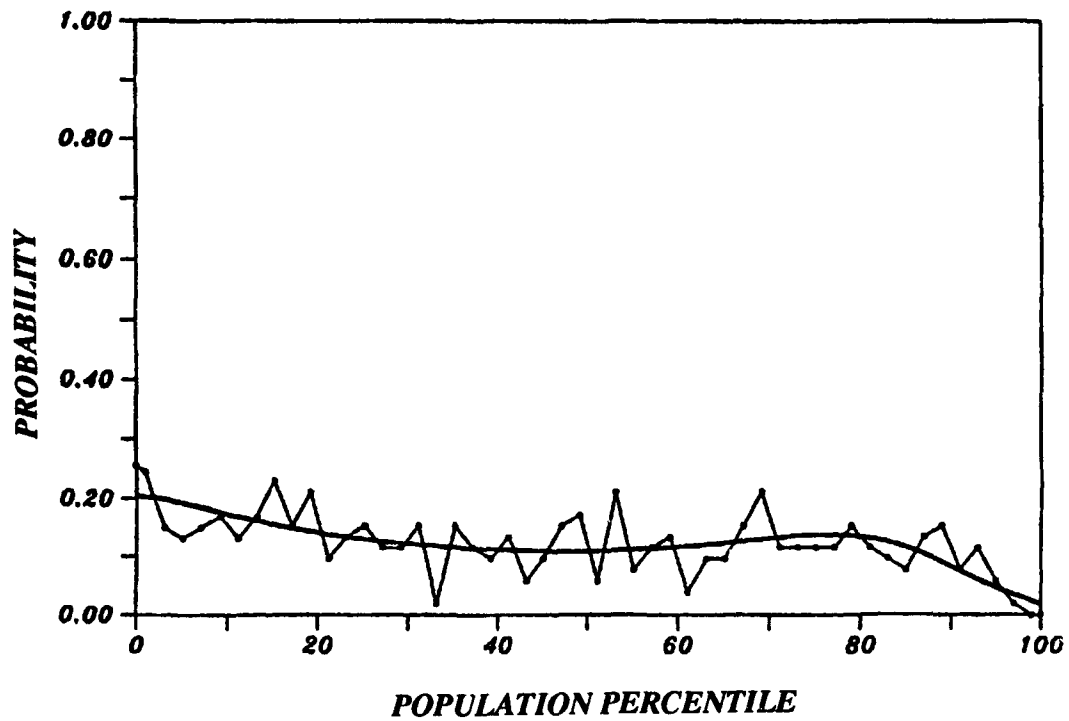**Figure 15. OCF for Item 6, Category 3, under Model 8.**

13

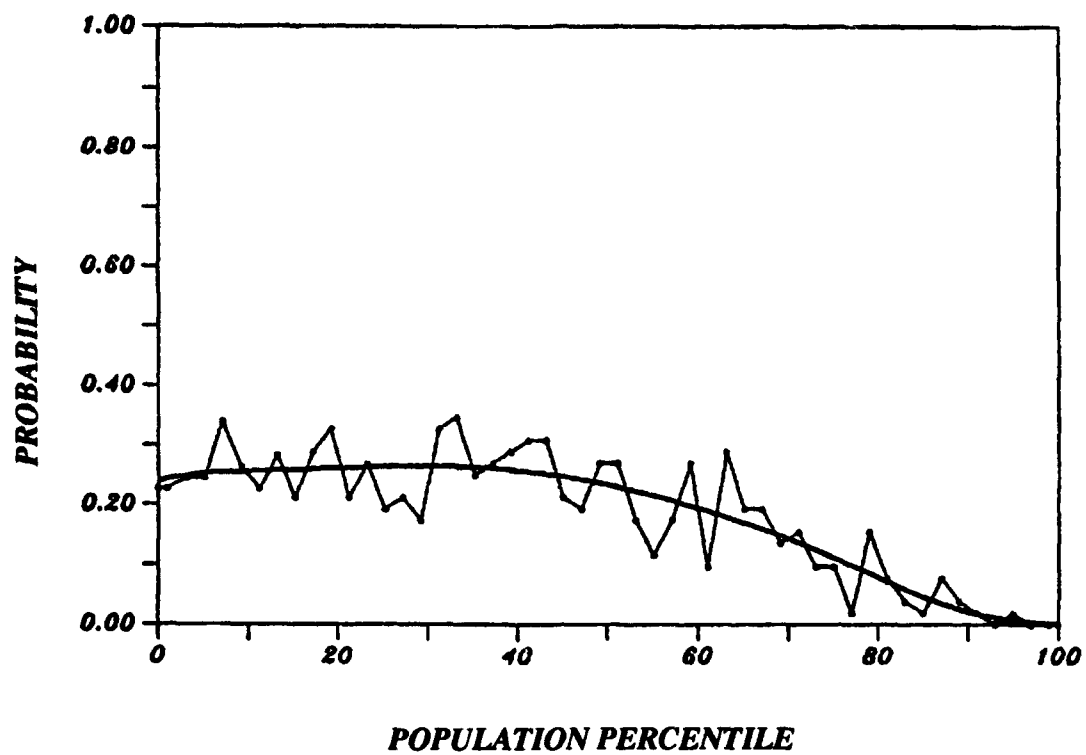Figure 16. OCF for Item 6, Category 4, under Model 8.
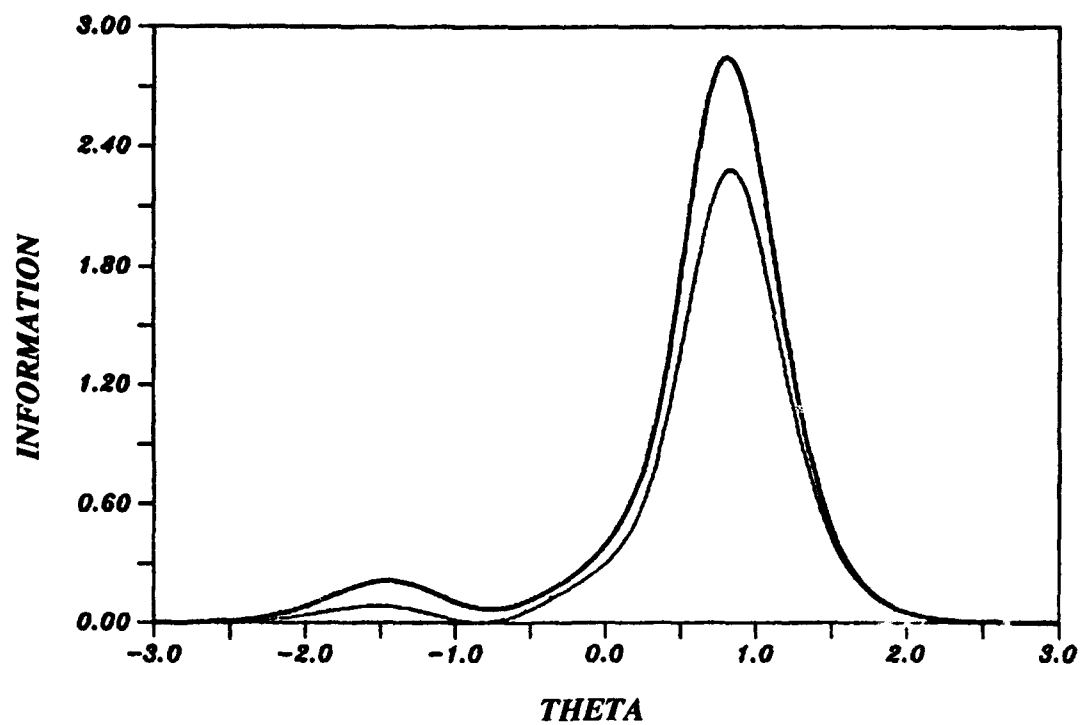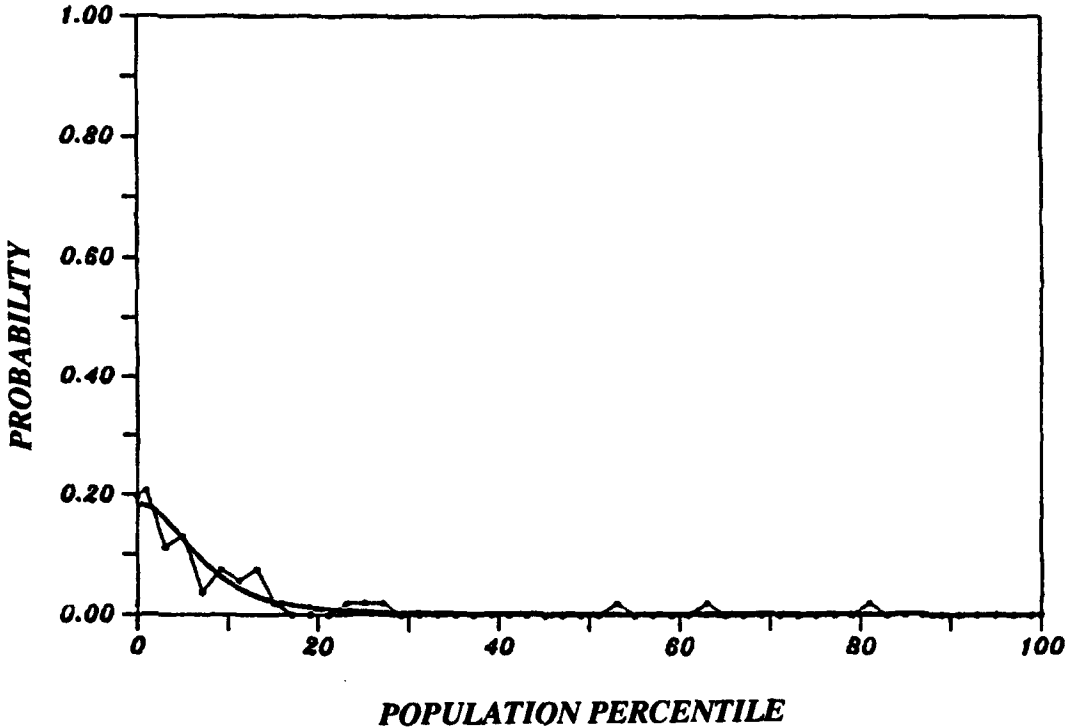

Figure 17. OCF for Item 6, Category 5, under Model 8.

14
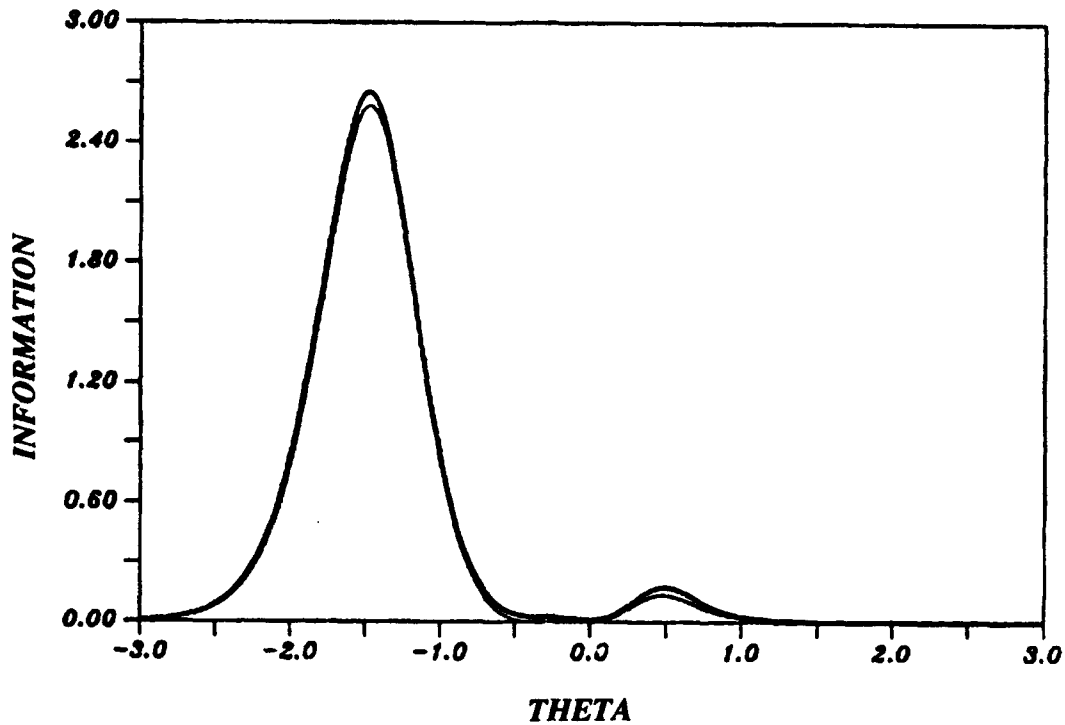
Figure 18. Item information functions for Item 6 under polychotomous and dichotomous scoring.

Figures 19 through 23 show the OCFs for Item 25, an item with one very popular distractor (Category 3, shown in Figure 21). This distractor was selected by over 60% of the examinees between the 15th and 75th percentiles of the military applicant sample. The impact of this distractor on the OCF for the correct-answer category (Figure 22) is obvious. The polychotomous item information function for this item (Figure 24) is strongly bimodal, which reflects the fact that Category 3 provides a substantial amount of additional information in the ability range where its OCF is rapidly increasing.

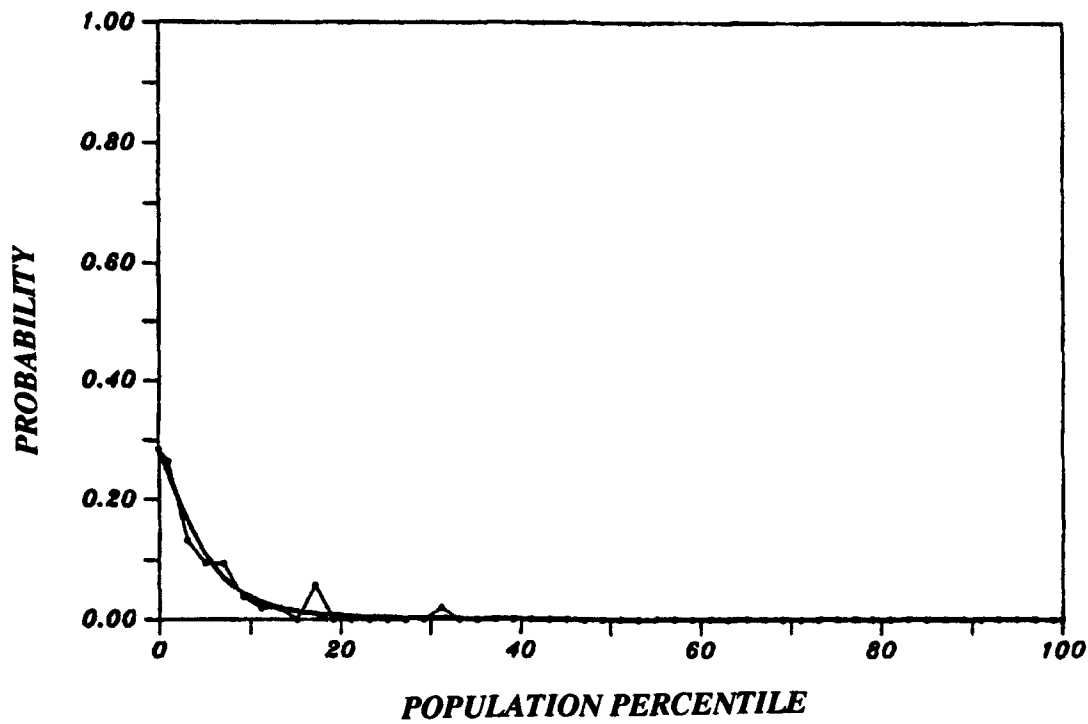**Figure 19. OCF for Item 25, Category 1, under Model 8.**

Figure 20. OCF for Item 25, Category 2, under Model 8.
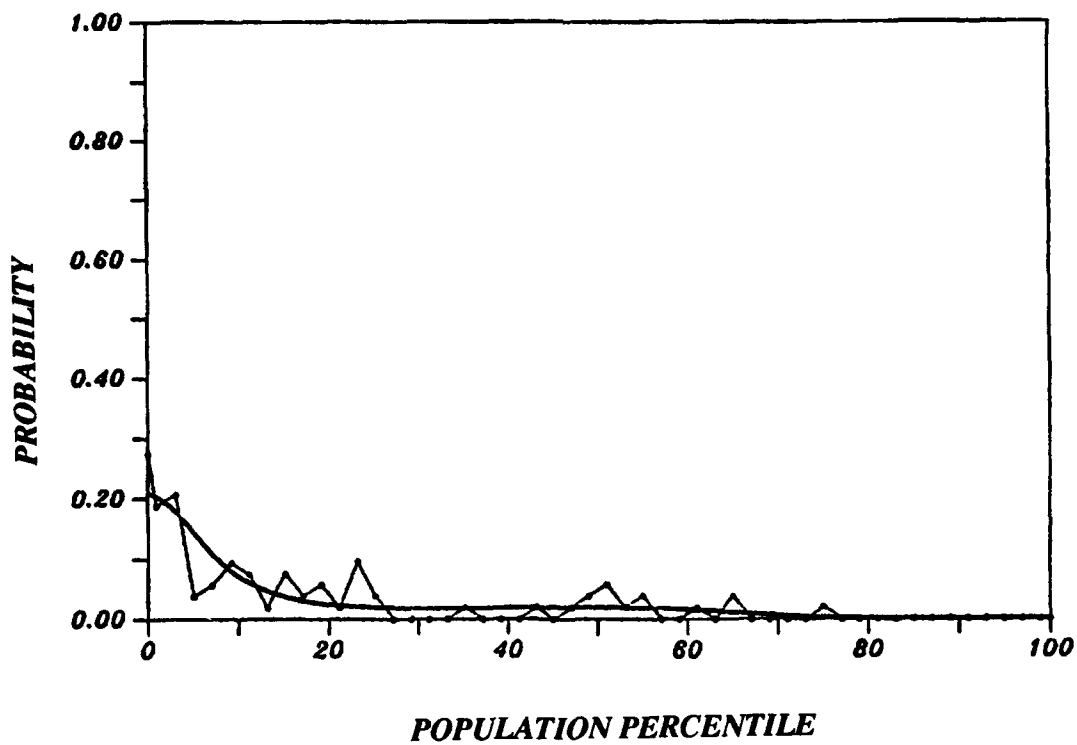
16

Figure 21. OCF for Item 25, Category 3, under Model 8.
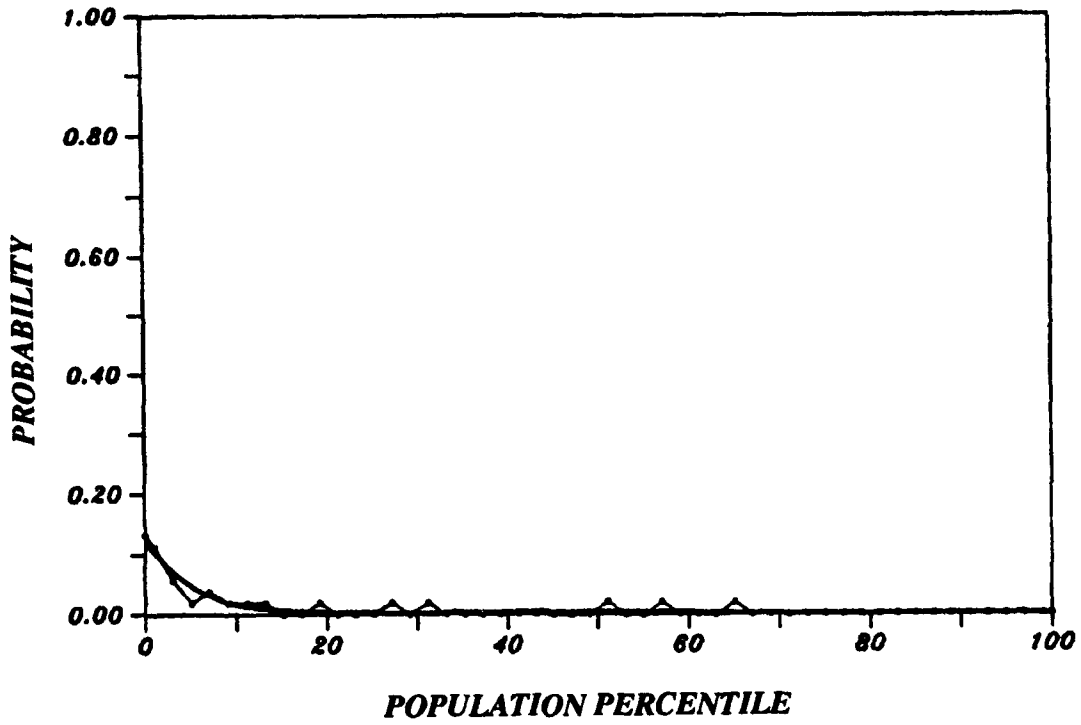

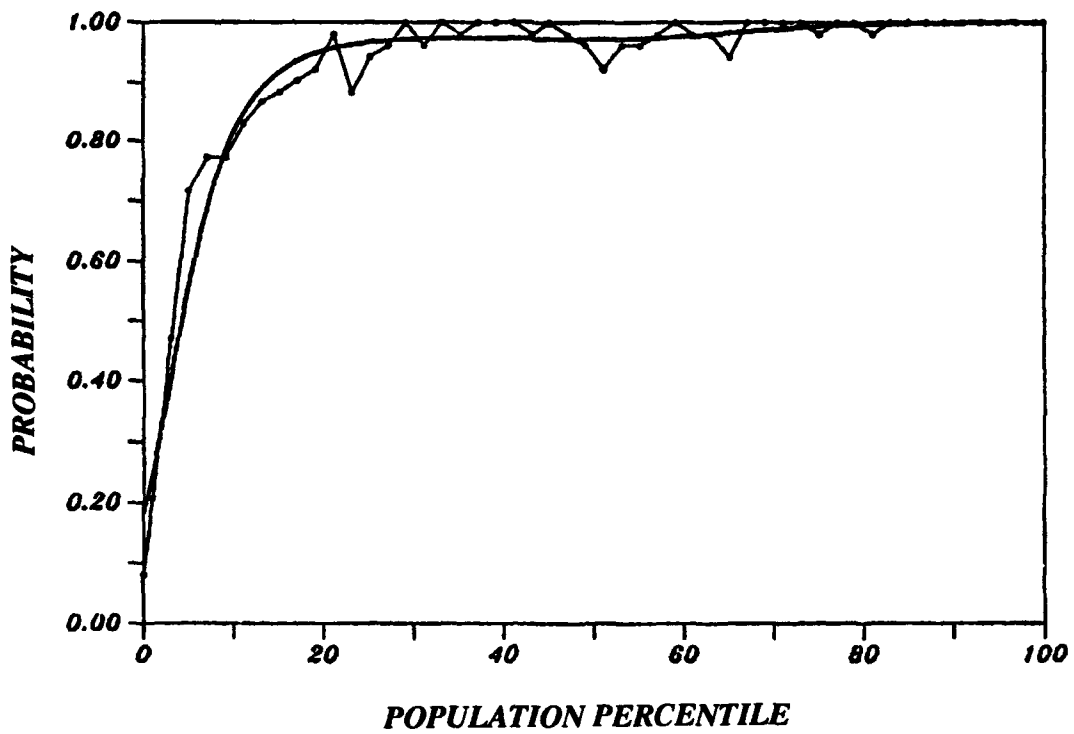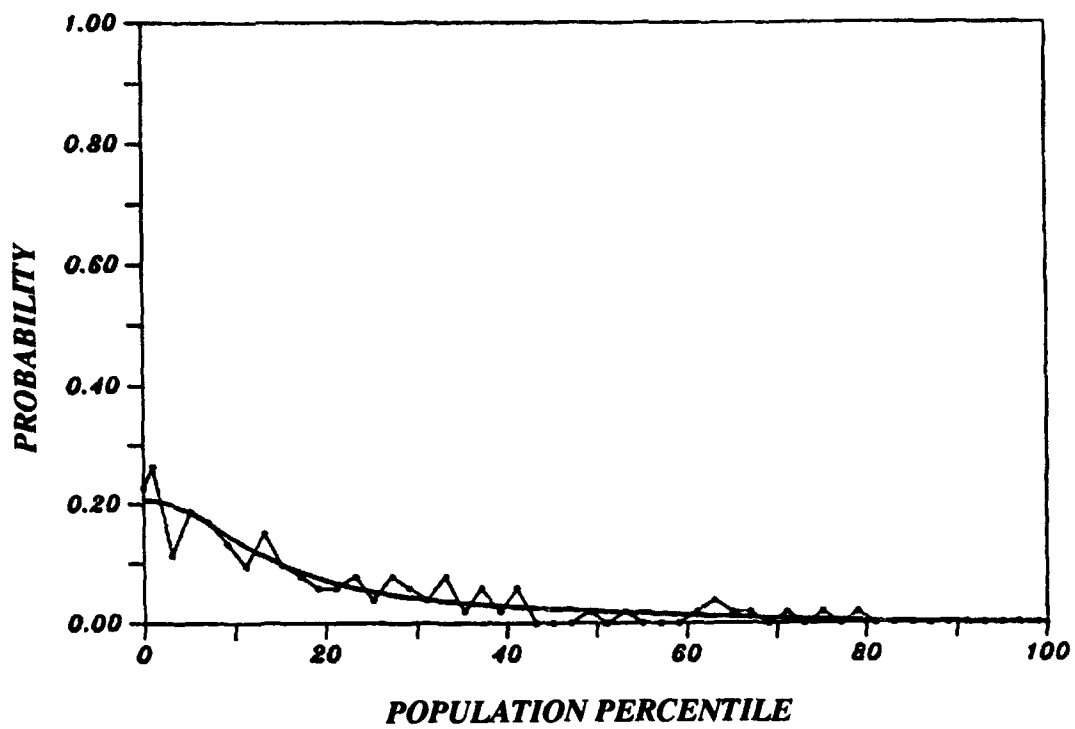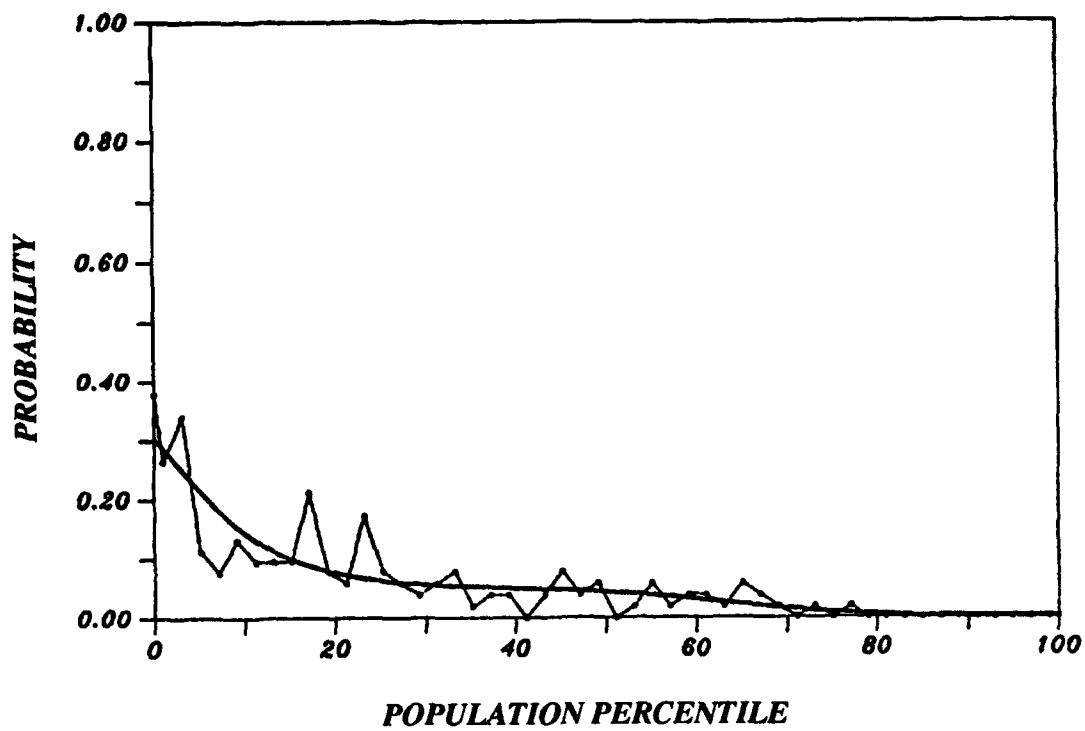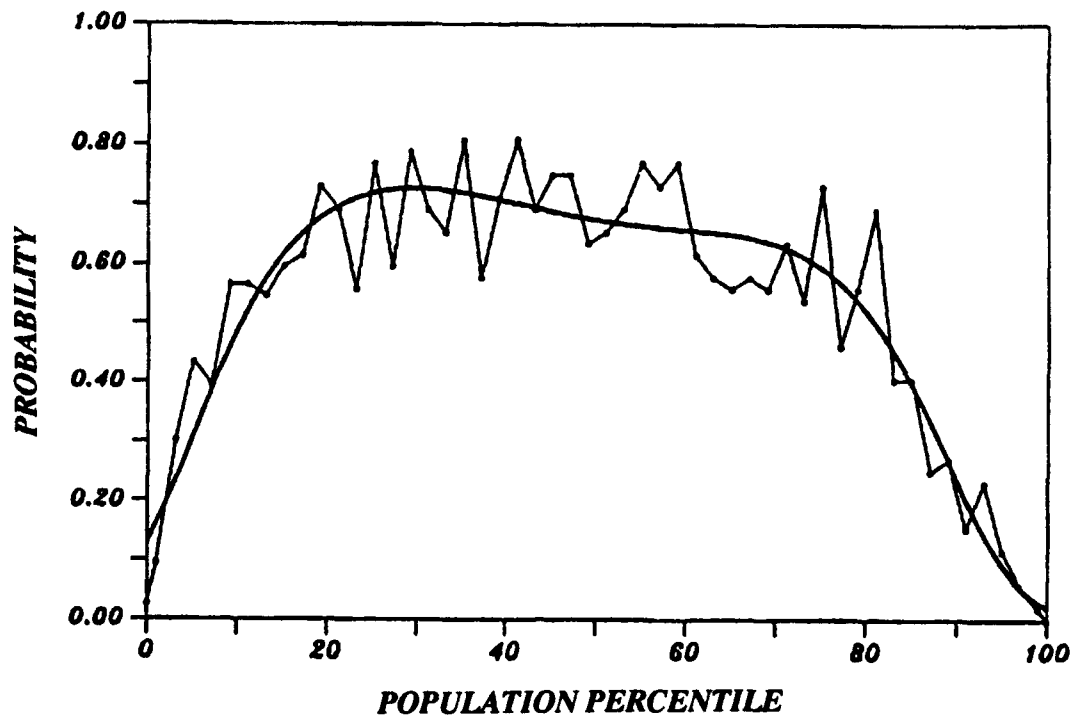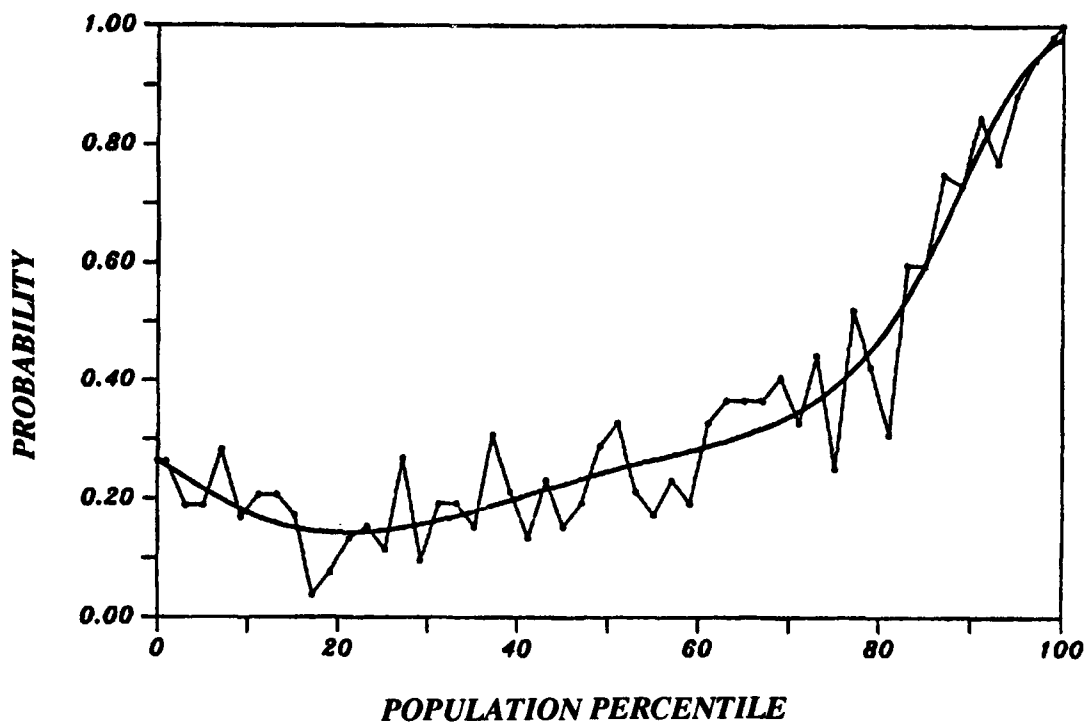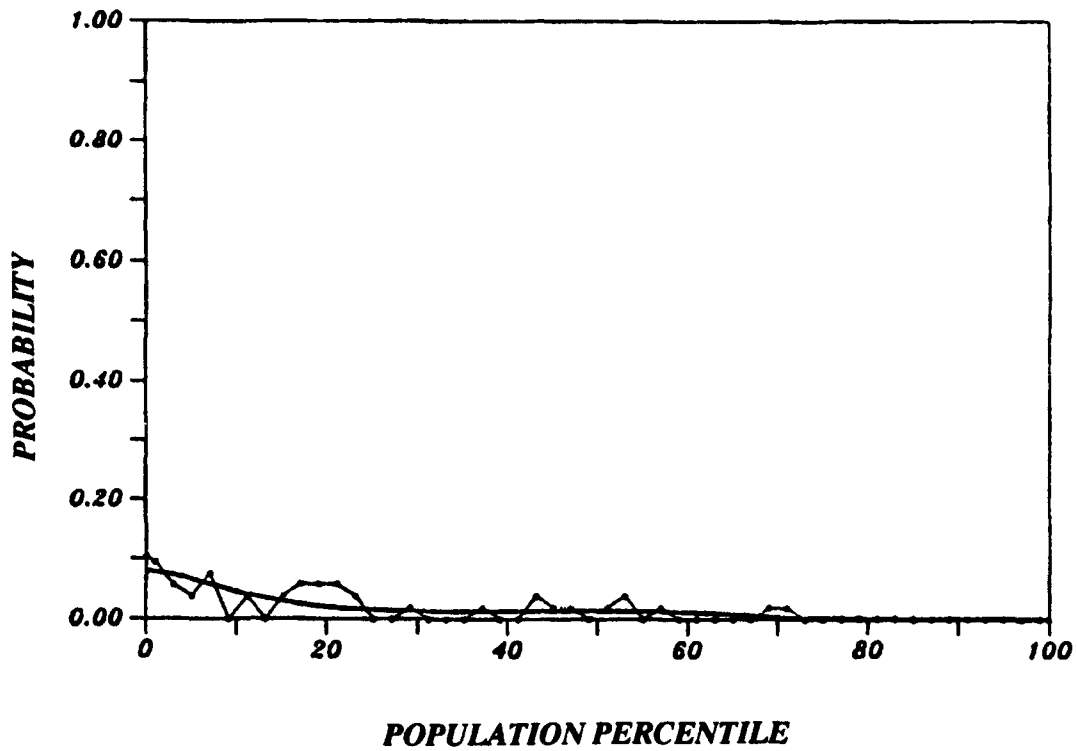**Figure 22. OCF for Item 25, Category 4, under Model 8.**

17

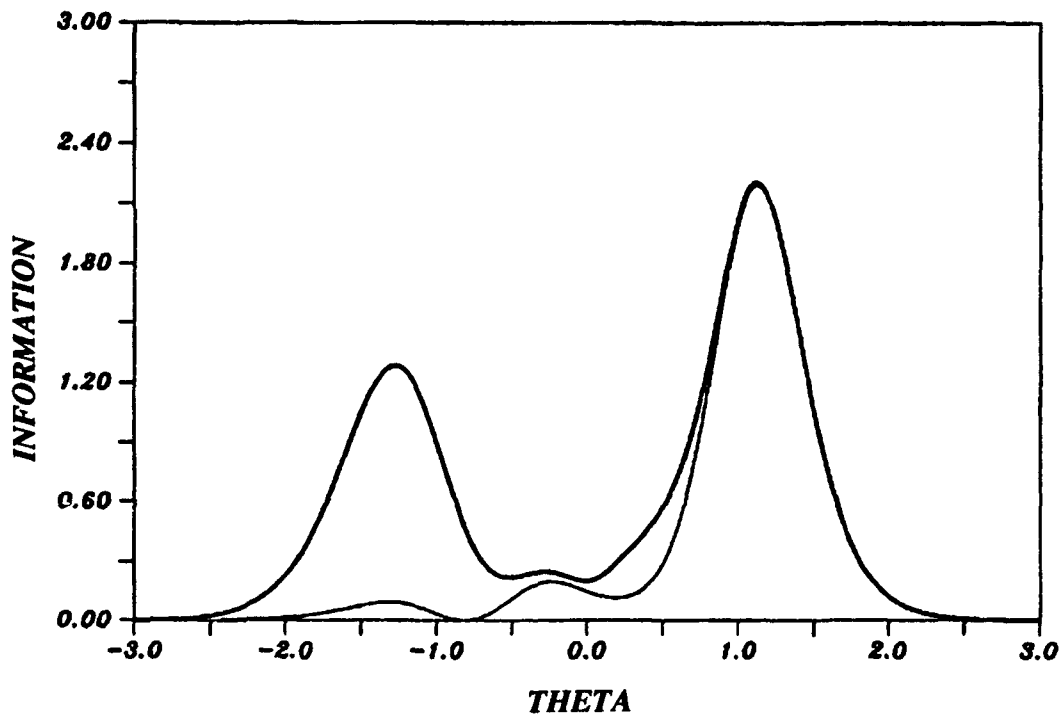**Figure 23. OCF for Item 25, Category 5, under Model 8.**

Figure 24. Item information functions for Item 25 under polychotomous and dichotomous scoring.

18

In general, careful inspection of the content of items that have one or more effective distractors reveals the source of the attractiveness of these distractors. As an example, Item 25 asks the examinee to select the best synonym for "wary." The keyed answer is "careful" (Category 4). Category 3, the very popular distractor, is "tired." It appears that many middle-ability military applicants confuse the words "wary" and "weary."

It might be argued that Item 25 is a bad item, that it measures spelling ability rather than vocabulary skills. An attempt to resolve that issue will be not be presented here. However, two observations related to the issue should be noted. First, examinees in the top 5% of the sample do not tend to confuse "wary" with "weary." These individuals are not just good spellers. They have demonstrated, by their performance on the entire set of 86 items, that they have superior knowledge of word meanings. Second, *every* item in a printed vocabulary test is, to some extent, a "spelling" item.

The OCFs shown and the discussion of Item 25 point up two benefits of using Model 8 for item calibration. First, this model can fit items that would be fitted poorly by simpler models. Model 8 allows non-monotone OCFs for correct-answer categories and fits lower (and upper) asymptotes separately for each response category. Second, using this type of model directs the user's attention to the psychological processes underlying examinee responses. This can be quite enlightening when one is interested in conducting individual diagnostic evaluations of persons and/or test items.

Method

Adaptive Testing With Model 8

The fact that Model 8 sometimes fits correct-answer categories better than the 3PL model does, and the fact that many item information functions show higher levels of information under polychotomous item scoring than under dichotomous scoring, suggest that test scores derived using Model 8 should be more reliable than scores derived using the 3PL model. In particular, this should be true for conventional tests, where item difficulty is not tailored to the individual examinee and wrong answers are frequent at low ability levels.

The outcome may be less clear-cut if we apply Model 8 in the context of CAT. In CAT, the difficulty of test questions is matched to the estimated ability level of the examinee. As a result, the number of incorrect answers is controlled by the testing procedure. CAT procedures usually result in about 70% correct answers and 30% incorrect answers for all examinees except those at the extremes of the ability distribution.

In order to investigate the application of Model 8 to CAT, the following analysis was undertaken. First, the 86 vocabulary items that had been calibrated with both the 3PL model and *Model 8* were divided into two 43-item sets, based on their sequence numbers within the original test booklet. Odd-numbered items were assigned to one set and even-numbered items to the other. Then, two "information tables" were constructed for each 43-item set. For each item set, one information table was based on item information functions derived from the items' fitted 3PL OCFs (Birnbaum, 1968, p. 462), and the other information table was based on item information functions derived from the items' fitted Model 8 OCFs.

An information table is constructed by selecting closely-spaced θ levels over a specified range of θ, and then sorting the available items from "best" to "worst" in terms of the value of the item information function at each selected θ level. For the purposes of this analysis, θ levels ranging from -2.50 to 2.50 in .10 steps were used in constructing both of the "odd" information tables and both of the "even" information tables. Thus, each of the four information tables had 51 columns (corresponding to the selected θ levels) and 43 rows. Each column of an information table contained 43 item numbers (either odd or even), sorted so that the item numbers of items with higher levels of item information at that θ level were located above the item numbers of items with lower levels of item information.

The purpose of constructing an information table is to provide a rapid procedure for selecting items to administer during CAT. In practice, one usually starts an adaptive test by assuming the examinee has ability equal to the population mean (e.g., θ = 0.0). For this analysis, the item appearing at the top of a given information table in the column corresponding to θ = 0.0 was administered first. Then, an examinee's response was noted and the initial ability estimate was modified in accordance with that response. The modified ability estimate was then used to identify the column of the information table that should be used next. The most informative item available in that column of the table was administered second. When an item was administered, it was immediately removed from *all* columns of the information table for the remainder of that examinee's test.

The process of selecting an item from the information table, modifying the ability estimate in accordance with the examinee's response, and moving to the column of the table that corresponded to the revised ability estimate was repeated until 15 items had been administered.

For the analyses undertaken here, two methods of generating ability estimates were used. One method (developed by Owen, 1975), was applied when the information tables based on the 3PL model were used. The other method, based on a numerical solution to the equation

$$\mu(\theta \mid V_n) = [\int \{L(V_n \mid \theta) \; h(\theta)\} \; d\theta]^{-1} \; [\int \{L(V_n \mid \theta) \; h(\theta) \; \theta\} \; d\theta] \tag{4}$$

was applied when the information tables based on Model 8 were used.

In Equation 4, V_n designates the observed item response vector after n items have been administered, $\mu(\theta \mid V_n)$ is the mean of the Bayesian posterior distribution of θ, given V_n, $L(V_n \mid \theta)$ is the likelihood of V_n, which is obtained by taking the product over items of $P_{ij}[\pi(\theta)]$, and $h(\theta)$ is the prior density of θ in the examinee population. Sympson (1985, pp. 3-6) discusses the pros and cons of these two approaches to Bayesian ability estimation.

Owen's ability estimation procedure was developed explicitly for use with the 3-parameter normal-ogive model, but can be used with the 3PL model because of the similarity of the OCFs for these two models. Owen's procedure cannot be used with a polychotomous model like Model 8. The second ability estimation method, which requires numerical integrations to evaluate the right side of Equation 4, is generic and can be used with any model.

To carry out the numerical integrations needed for the second ability estimation method, an "adaptive" quadrature procedure described by Forsythe, Malcolm, and Moler (1977, pp. 97-105)

20

was employed. Numerical integration using fixed-point quadrature to obtain Bayesian ability estimates has been used previously with the 3PL model by Sympson (1977) and Bock and Mislevy (1982), but, to the present author's knowledge, this study is the first to use an adaptive quadrature procedure and is the first to implement Bayesian ability estimation with a polychotomous item-response model.

In the analyses undertaken here, each of the 2,607 examinees used in calibrating the 86 vocabulary items was "administered" four simulated adaptive tests: Test "Odd-3PL" used the odd-item 3PL information table and Owen's scoring method with the fitted 3PL OCFs; Test "Evn-3PL" used the even-item 3PL information table and Owen's scoring method with the fitted 3PL OCFs; Test "Odd-M8" used the odd-item Model 8 information table and numerical scoring with $L(V_n \mid \theta)$ computed from the fitted Model 8 OCFs; Test "Evn-M8" used the even-item Model 8 information table and numerical scoring with $L(V_n \mid \theta)$ computed from the fitted Model 8 OCFs. In all four tests, the prior distribution of θ was assumed to be normal (0,1).

In each of the simulated adaptive tests, an examinee's actual item responses from the empirical item-calibration database were used. Within each test, an ability estimate was generated after each selected item was "administered". After all tests were completed, Pearson product-moment correlations were computed between the 2,607 "odd" and "even" 3PL ability estimates. This was done for the ability estimates based on the first item administered, the ability estimates based on the first two items administered, and so on, up to the final ability estimates based on all 15 adaptively selected items. Similar "odd-even" correlations were computed using the ability estimates obtained from tests Odd-M8 and Evn-M8. The outcome of these analyses was a set of 15 odd-even correlations based on the 3PL model and a set of 15 odd-even correlations based on Model 8. These correlation coefficients are shown in Table 1.

Results and Discussion

The results shown in Table 1 are not entirely consistent with theoretical expectations. Theory suggests that the Model 8 correlation should be higher at all test lengths, with the difference diminishing as the test becomes very long. However, the odd-even correlation for the 3PL model is somewhat higher than the corresponding correlation for Model 8 after Items 2, 3, and 4. Correlations under the two models are approximately equal after Items 5 and 6. The odd-even correlation under Model 8 is higher after Item 1 and after Items 7 through 15. As indicated in the first column of Table 1, differences between odd-even correlations are statistically significant following Items 9 through 15.

The odd-even correlations obtained under Model 8 after Items 9 and 10 have been administered are higher than the odd-even correlations obtained after administering one item more under the 3PL model (e.g., $r = .855$ after 9 items for Model 8 versus $r = .852$ after 10 items for 3PL). The odd-even correlation obtained under Model 8 after 11 items is higher than the correlation obtained after administering 13 items under the 3PL model. The odd-even correlation after 12 items under Model 8 is approximately equal to the final (15-item) odd-even correlation under the 3PL model. These results indicate that if Model 8 were used, adaptive tests of the type studied here could be shortened by 15 to 20%, without sacrificing test reliability.

21

Table 1

Odd-Even Correlations After *n* Items for the
3PL Model and for Model 8

	Odd-Even Correlations	
n	3PL Model	Model 8
1	.363	.366
2	.544	.536
3	.661	.651
4	.721	.709
5	.767	.766
6	.795	.796
7	.817	.823
8	.832	.840
9*	.843	.855
10**	.852	.865
11**	.860	.875
12**	.869	.881
13**	.873	.887
14**	.877	.891
15**	.880	.895

*Correlations significantly different ($p < .01$).
**Correlations significantly different ($p < .001$).

There is no doubt that Model 8 provided a better fit to the correct-answer probabilities of some of the 86 items studied than did the 3PL model. Inspection of goodness-of-fit plots for both models (examples were shown above for Model 8) made this clear. Also, for many items the item information function was higher under polychotomous scoring than under dichotomous scoring. These advantages translated into higher odd-even correlations for Model 8 at test lengths of 7 or more items. Further investigation is needed to determine the reason that odd-even correlations were higher under the 3PL model very early in the adaptive tests.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Department of Defense (1984). *Test Manual for the Armed Services Vocational Aptitude Battery* (DoD 1304.12AA). North Chicago, IL: U.S. Military Entrance Processing Command.

Dixon, W. J. (Ed.). (1981). *BMDP statistical software 1981*. Berkeley, CA: University of California Press.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1977). *Computer methods for mathematical computations*. Englewood Cliffs, NJ: Prentice-Hall.

Hambleton, R. K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 71-94). Vancouver, BC: Educational Research Institute of British Columbia.

Jennrich, R. I., & Moore, R. H. (1975). *Maximum likelihood estimation by means of nonlinear least squares* (RB-75-7). Princeton, NJ: Educational Testing Service.

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

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4*, 321-334.

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Research Report 79-4). Knoxville, TN: University of Tennessee, Department of Psychology.

Strang, H. R. (1977). The effects of technical and unfamiliar options on guessing on multiple-choice test items. *Journal of Educational Measurement, 14*, 253-260.

Sympson, J. B. (1977). Estimation of latent trait status in adaptive testing procedures. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing* (Research Report 77-1, pp. 5-23). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Sympson, J. B. (1981, October). *A nominal model for IRT item calibration*. Talk given at the Office of Naval Research Conference on Model-based Psychological Measurement, Millington, TN.

Sympson, J. B. (1983, June). *A new item response theory model for calibrating multiple-choice items*. Paper presented at the meeting of the Psychometric Society, Los Angeles, CA.

Sympson, J. B. (1984, October). *Principal component analysis of polychotomous item responses*. Talk given at the Office of Naval Research Conference on Model-based Psychological Measurement, Princeton, NJ.

Sympson, J. B. (1985, June). *Bayesian estimation of true scores and observed scores on a criterion test*. Paper presented at the meeting of the Psychometric Society, Nashville, TN.

Sympson, J. B. (1986a). Models for calibrating multiple-choice items. In R. Penn & A. Crawford (Eds.), *Independent Research and Independent Exploratory Development FY85* (NPRDC Special Report 86-1, pp. 1-4). San Diego: Navy Personnel Research and Development Center.

Sympson, J. B. (1986b, April). *Some item response functions obtained in polychotomous item analysis*. Talk given at the Office of Naval Research Conference on Model-based Psychological Measurement, Gatlinburg, TN.

Sympson, J. B. (1986c, August). Extracting information from wrong answers in computerized adaptive testing. Paper presented in B. F. Green (Chair), *New developments in computerized adaptive testing*. Symposium conducted at the annual meeting of the American Psychological Association, Washington, DC.

Thissen, D., & Steinberg, L. (1983). *A response model for multiple choice items* (Psychometric Technical Report No. 1). Chicago: University of Chicago, National Opinion Research Center.

Distribution List

Distribution:
Office of the Assistant Secretary of Defense (FM&P)
Office of Naval Research (Code 1142) (3)
Defense Technical Information Center (DTIC) (12)

Copy to:
Office of Naval Research (Code 20P), (Code 222), (Code 10)
Naval Training Systems Center, Technical Library (5)
Office of Naval Research, London
Director, Naval Reserve Officers Training Corps Division (Code N1)
Chief of Naval Education and Training (L01) (2)
Curriculum and Instructional Standards Office, Fleet Training Center, Norfolk, VA
Chief of Naval Operations (N71)
Director, Recruiting and Retention Programs Division (PERS-23)
Commanding Officer, Sea-Based Weapons and Advanced Tactics School, Pacific
Commanding Officer, Naval Health Sciences Education and Training Command, Bethesda, MD
Marine Corps Research, Development, and Acquisition Command (MCRDAC), Quantico, VA
AISTA (PERI II), ARI
Armstrong Laboratory, Human Resources Directorate (AL/HR), Brooks AFB, TX
Armstrong Laboratory, Human Resources Directorate (AL/HRMIM), Brooks AFB, TX
Armstrong Laboratory AL/HR-DOKL Technical Library, Brooks, AFB, TX
Library, Coast Guard Headquarters
Superintendent, Naval Post Graduate School
Director of Research, U.S. Naval Academy
Naval Education and Training Program (NETPMSA, Code 047), Pensacola (N. N. Perry)