AFHRL-TP-89-15

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

AD-A216 121

APPROPRIATENESS MEASUREMENT FOR COMPUTERIZED ADAPTIVE TESTS

Gregory L. Candell
Michael V. Levine

Model Based Measurement Laboratory
210 Education Building
University of Illinois
1310 South Sixth Street
Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

December 1989
Final Technical Paper for Period October 1987 - June 1989

ELECTE
DEC 28 1989
S B D

Approved for public release: distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

89 12 28 013

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

DANIEL L. LEIGHTON, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

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	3. REPORT TYPE AND DATES COVERED
	December 1989	Final -- October 1987 to June 1989

4. TITLE AND SUBTITLE

Appropriateness Measurement for Computerized Adaptive Tests

5. FUNDING NUMBERS

C - F41689-87-D-0012
PE - 62730F
PR - 2922
TA - 02
WU - 02

6. AUTHOR(S)

Gregory L. Candell
Michael V. Levine

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Model Based Measurement Laboratory
210 Education Building
University of Illinois
1310 South Sixth Street
Champaign, Illinois 61820

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFHRL-TP-89-15

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The effects of an initial sequence of random responses to 15-, 20-, and 25-item adaptive tests were examined in a series of simulation studies. Random responding on as few as two items had a substantial effect on an examinee's score. Thus it is important to determine whether--as a result of carelessness, test anxiety, computer anxiety, failure to understand instructions, or other reasons--an examinee has answered the first several items haphazardly.

It was shown by use of an optimal appropriateness index, the likelihood ratio (LR) index, that a large proportion of faulty test scores can be identified. The performance of LR was evaluated by determining hit rates and false positive rates in a series of studies concerning: (a) comparisons with other indices, (b) the use of a security procedure during item selection for the adaptive test, (c) standardization, and (d) misspecification of the number of items with random answers.

The LR index detected initial sequences of random responses with high accuracy with and without a security procedure during item selection. Other appropriateness indices were considerably less accurate. Standardization greatly decreased the power of LR at low false positive rates. Finally, misspecification of the length of the initial segment of random responses systematically reduced the power of the LR index to detect aberrance.

14. SUBJECT TERMS

appropriateness measurement
aptitude tests
computerized adaptive test

15. NUMBER OF PAGES

64

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL Technical Paper 89-15 December 1989

APPROPRIATENESS MEASUREMENT FOR COMPUTERIZED ADAPTIVE TESTS

Gregory L. Candell
Michael V. Levine

Model Based Measurement Laboratory
210 Education Building
University of Illinois
1310 South Sixth Street
Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed b

Linda T. Curran, Actin[illegible] hief
Enlisted Selection and Classification Function

Submitted for publication by

Lonnie D. Valentine, Jr., Chief
Force Acquisition Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

There is an ongoing effort to computerize the Armed Services Vocational Aptitude Battery (ASVAB), the enlisted selection and classification test, into an adaptive mode. In an adaptive test, the computer estimates how well the examinee is doing after each item has been answered and then selects the next item--more difficult for brighter examinees or easier for less bright examinees--from a large pool of items in its memory bank. Modern test theory allows examinees to be appropriately scored even though they have each been administered completely different sets of items. This report demonstrates that examinees who are confused by the computerized administration medium and give inappropriate responses to just a few of the initial items (i.e., give responses that are not representative of their actual knowledge) will be severely penalized in their final score. Applications of appropriateness measurement techniques show that it is possible to identify such examinees with highly computer-intensive calculations. Use of short-cut formulas which have been found useful under some conditions for appropriateness measurement give results that are much less optimal. Issues for further research are discussed.

Accession For	
NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

PREFACE

This technical paper was completed as part of the research conducted under Work Unit 29220202, Prototype Development and Validation of Selection and Classification Instruments, under Project 2922, Personnel Assessment Systems.

The research on appropriateness for computer adaptive tests presented in this paper is ancillary to Air Force responsibility on appropriateness of examinee responses on the Armed Services Vocational Aptitude Battery.

The Air Force acknowledges the Army for its initial guidance and funding support on this effort.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPROPRIATENESS MEASUREMENT FOR COMPUTERIZED ADAPTIVE TESTS

I. INTRODUCTION

Computerized adaptive testing (CAT) and appropriateness measurement (Levine & Drasgow, 1982; Levine & Rubin, 1979) are two promising applications of item response theory (IRT). The potential benefits of each application represent important advances in testing and measurement. CAT offers advantages over conventional, paper-and-pencil tests such as reduced test length and equivalent measurement precision across the range of test scores (Weiss, 1982). Appropriateness indices provide the capability of detecting spurious test scores that result from situations such as cheating or alignment errors in marking answer sheets.

To date, appropriateness measurement has been applied to conventional tests only. This paper presents research that examines the potential for appropriateness measurement in adaptive testing. More specifically, methods from appropriateness measurement will be used to attempt to detect one type of response aberrance that may have a serious impact on some CAT scores.

In the remainder of this introduction, CAT and appropriateness measurement will be reviewed briefly, a potential form of response aberrance for CAT will be identified, and the scope of the present research will be outlined.

Computerized Adaptive Tests. CAT is designed to administer the set of items, from a larger pool of items, that provide optimal measurement of each examinee's ability. This is accomplished by tailoring the difficulty of each item administered to the current ability estimate (as calculated from responses to the preceding items). In CAT, items that are too easy or too difficult to provide information about ability are not administered; instead, examinees receive items during the test that are highly informative of their ability.

The ability parameter and item difficulty parameter share the same scale in IRT, making it possible in CAT to select and administer items of appropriate difficulty. Although a variety of item selection strategies

exist for CAT, each of these strategies is consistent with the logic of matching item difficulty with estimated ability.

The potential of reduced test length for CAT requires that items perform in accordance with their parameter estimates (Wainer & Kiely, 1987). More generally, the IRT model used by the test must be able to account for an examinee's response pattern if the measurement virtues of CAT are to be realized. Random departures from the model can be expected for some examinees on some items. However, the impact of occasional misinformative responses on the final CAT ability estimate is generally believed to be small, especially if test termination is based on reducing the estimate's standard error to an acceptable level and a sufficient number of items exist to accommodate this criterion.

Appropriateness Measurement. For a variety of reasons, a multiple-choice test may fail to provide a valid measure of ability for an examinee. For example, answers may be copied from a more talented neighbor, resulting in a spuriously high test score. Spuriously low scores may result from circumstances such as alignment errors in marking the answer sheet (e.g., answering items 5 through 10 in the spaces provided for items 6 through 11), cultural and/or linguistic bias, or extreme test anxiety. In each of these cases, the test score may be an invalid measure of the trait purportedly measured by the test.

Appropriateness measurement provides several IRT-based methods for identifying such scores. Each method develops a quantitative index to classify item response patterns as either "normal" or "aberrant." Appropriateness measurement is model-based; normal response patterns are characterized as conforming to a specific IRT model for describing item responses. In this sense, appropriateness indices are goodness-of-fit tests for response patterns relative to an IRT model. The logic that underlies aberrance detection for dichotomously scored, unidimensional tests is straightforward: Inappropriate response patterns will contain correct responses to difficult items co-occurring with incorrect responses to easy items.

Simulation studies using appropriateness indices to classify aberrant and normal response patterns have obtained high levels of detection for some forms of aberrance on standardized tests (e.g., Drasgow, Levine, &

McLaughlin, 1987; Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979). High detection rates have been achieved despite misspecification of the IRT model, errors in item parameter estimates, and inclusion of inappropriate response patterns in the test norming sample (Levine & Drasgow, 1982).

The properties of CAT may allow for successful applications of appropriateness measurement. Since the difficulties of administered items are determined by examinee responses and the IRT model, some types of aberrance may be easier to identify than with standardized tests.

On the other hand, the relatively short length of adaptive tests may not provide sufficient numbers of items to powerfully test whether a response pattern departs from the pattern expected under a given IRT model. Molenaar and Hoijtink (1987) have noted that since adaptive tests are relatively homogeneous with respect to item difficulty, there may not be sufficient variance among item difficulties to detect inappropriate patterns of response. They recommend that appropriateness measurement not be applied to tests of less than 20 items, since random fluctuations may dominate systematic departures from the IRT model over small numbers of items.

CAT Response Aberrance. Though applications such as CAT have generated interest among measurement practitioners, several concerns for computerized testing have been raised (Hunt & Pellegrino, 1985; Matarazzo, 1983). One concern is that taking the test on a computer may present a significant advantage or disadvantage to some examinees. For example, examinees with little or no previous computer experience may initially be intimidated by the task of taking a computerized test. Test performance may suffer as a result. The negative impact of anxiety on attention and cue utilization in tasks such as psychological tests is well documented (Broadbent, 1971; Easterbrook, 1959; Kahneman, 1973).

For CAT examinees, computer/test anxiety could result in test-taking behavior that departs significantly from the IRT model used by the test. For these examinees, item responses may appear to be the product of a random process rather than the function of item and person parameters prescribed by the IRT model. Anxious CAT examinees might experience difficulty in focusing attention on items throughout the test. Some of these anxious examinees, however, might "settle down" at some point during the test and respond to the remaining items in a manner consistent with the IRT model.

The possibility of this latter situation raises an important question for CAT: If an examinee "fumbles" through initial items and then "recovers" to respond more appropriately throughout the remainder of the test, does the adaptive test recover as well and provide accurate results?

Theoretically, adaptive tests are robust to sequences of misinformative responses. If enough items are administered, the impact of these aberrant responses can presumably be minimized and accurate ability estimates may be obtained. In practical CAT settings this may not hold true, however. The algorithms used for item selection and ability estimation may not be robust to sequences of misinformative responses, particularly when they occur at the outset of the test. Short, fixed-length adaptive tests, such as those planned for use with the Armed Services Vocational Aptitude Battery (ASVAB), may be particularly vulnerable.

Present Research. The present research contains a series of Monte Carlo studies designed primarily to address three issues: (a) the impact of an initial sequence of random responses on CAT results, (b) the potential of appropriateness measurement for detecting this type of response aberrance, and (c) the performance of the optimal appropriateness index under practical testing conditions such as standardizing the index and using an item security procedure to administer CAT items.

Study 1 examines the robustness of short, fixed-length adaptive tests to an initial sequence of random responses. Study 2 focuses on the highest possible detection rates for this form of response aberrance on a 15-item adaptive test. Results from Study 2 will determine whether any appropriateness index computed from dichotomous responses can provide the level of detection that practical testing situations will require.

Studies 3 through 6 examine several issues related to implementating appropriateness measurement for detecting random responses to initial CAT items. In Study 3, several nonoptimal appropriateness indices are examined to see if the performance of any of these indices approaches the optimal levels observed in Study 2. The sampling behavior of appropriateness measurement results using Monte Carlo procedures is addressed in Study 4. The results of Study 4 are then used in the designs for both Study 5 and Study 6, which examine, respectively, the effect of an item security procedure on detection rates and the generalizability of aberrance detection

using the optimal index under nonoptimal conditions. Study 6 investigates a standardized version of the optimal index initially used in Study 2 and the effect of standardization on its detection rates.

II. STUDY 1: ROBUSTNESS OF SHORT ADAPTIVE TESTS

Purpose. A sequence of misinformative responses to initial CAT items may have a serious impact on test results. This study assesses the effect of this type of response aberrance on CAT results for examinees of varying ability levels.

Data Generation. Ten ability levels corresponding to the 5th, 15th, ..., 85th, and 95th percentiles for a normal distribution were used in generating responses to a 15-item adaptive test. The item pool consisted of 100 items from the CAT-ASVAB Word Knowledge item pool. Expected *a posteriori* (EAP) (Bock & Mislevy, 1982) estimation was used to obtain an ability estimate after each item response and items were selected to maximize information at $\hat{\theta}$.

In the control (no aberrance) condition, dichotomous responses to each item were determined by calculating the probability of a correct response using the three-parameter logistic model (Birnbaum, 1968) and comparing this probability to a random number drawn from a uniform [0,1] distribution. If the response probability was greater than the random number, the item response was scored correct; otherwise, the response was scored as incorrect.

In the aberrance conditions, the initial one, two, three, four, or five responses for each simulated examinee were made to be random; for these items, random responding to a 5-option multiple-choice item was modeled by setting the probability of a correct response equal to .20. The remaining item responses were determined by calculating the response probability using the three-parameter logistic model.

Results. Table 1 shows the average $\hat{\theta}$s calculated over 1,000 simulees at each ability level and at each of the aberrance conditions. Results for the no aberrance condition ($k = 0$) show that the adaptive test tends to underestimate above-average abilities and overestimate below-average abilities, with this negative and positive bias becoming more pronounced at extreme ability levels. These results are consistent with other CAT research that has demonstrated the regression effect for Bayesian estimation (e.g., McBride, 1977; Weiss & McBride, 1984).

Table 1. Simulation Results for a 15-Item Adaptive Test When the Initial k Responses Are Random

	$k = 0$		$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
θ	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$
-1.640	-1.55	0.26	-1.53	0.25	-1.53	0.25	-1.56	0.25	-1.62	0.25	-1.68	0.26
-1.040	-0.99	0.25	-0.98	0.25	-1.03	0.25	-1.09	0.25	-1.22	0.25	-1.33	0.25
-0.670	-0.64	0.25	-0.68	0.24	-0.73	0.24	-0.85	0.25	-1.02	0.25	-1.14	0.25
-0.385	-0.39	0.24	-0.42	0.23	-0.51	0.24	-0.66	0.25	-0.88	0.25	-1.05	0.25
-0.125	-0.11	0.24	-0.19	0.23	-0.32	0.24	-0.51	0.25	-0.75	0.25	-0.99	0.25
0.125	0.13	0.24	0.01	0.24	-0.13	0.25	-0.39	0.26	-0.71	0.25	-0.95	0.25
0.385	0.36	0.24	0.21	0.24	0.02	0.25	-0.34	0.26	-0.66	0.26	-0.86	0.25
0.670	0.62	0.24	0.44	0.25	0.17	0.26	-0.21	0.26	-0.59	0.26	-0.90	0.25
1.040	0.98	0.23	0.76	0.27	0.32	0.27	-0.14	0.27	-0.57	0.26	-0.84	0.25
1.640	1.52	0.23	1.17	0.30	0.50	0.28	-0.09	0.28	-0.50	0.26	-0.83	0.25

Note: θ = Ability parameter used in generating normal item responses.

$\bar{\hat{\theta}}$ = Mean ability estimate.

$\overline{PSD}$ = Mean posterior standard deviation.

Comparisons of $\hat{\theta}$s and θs in Table 1 highlight the serious consequences of an initial sequence of misinformative responses on some CAT scores. The degree of measurement bias is quite severe in some cases. In general, these results show that the highest ability levels are associated with the largest levels of underestimation. Significant underestimation occurs for above-average abilities when the initial two item responses are misinformative, and this negative bias becomes increasingly pronounced as the number of misinformative responses increases. The most extreme example is shown in the case of the highest ability level (θ = 1.64): An examinee at the 95th percentile who gives random responses to the first five items of the adaptive test would obtain, on average, a test score below the 25th percentile. For less extreme cases, the problem is not as severe but is still significant. These results strongly suggest that a 15-item adaptive test, using well-accepted ability estimation and item selection strategies, does not recover from an initial sequence of random responses to yield accurate ability estimates.

Table 2 gives the results for adaptive tests of 20 and 25 items. For the 25-item test, the size of the pool of available items was increased to 200. Results for the control conditions for both tests show that the regression effect on $\hat{\theta}$s becomes less severe as additional items are administered. The results for the aberrant conditions, however, indicate that the effect of misinformative responses to initial items is not removed or significantly reduced by increasing test length. The level of underestimation observed for the 20-item adaptive test, after the initial five responses were random, is comparable to the results for the 15-item adaptive test where the initial four responses were random. The effect of an initial sequence of five random responses on ability estimates for a 25-item test is also significant. The results for the 25-item test are more severe than those obtained for the 15-item test when the initial three responses were random, but less severe than the results for the 15-item test when the initial four responses were random.

Table 2. Simulation Results for 20- and 25-Item Adaptive Tests When the Initial k Responses Are Random

	20-Item Test				25-Item Test			
	k = 0		k = 5		k = 0		k = 5	
θ	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$	$\bar{\hat{\theta}}$	$\overline{PSD}$
-1.640	-1.57	0.22	-1.67	0.22	-1.58	0.20	-1.68	0.20
-1.040	-1.00	0.22	-1.27	0.22	-1.00	0.20	-1.22	0.20
-0.670	-0.64	0.22	-1.03	0.22	-0.65	0.20	-0.99	0.20
-0.385	-0.37	0.21	-0.90	0.22	-0.38	0.19	-0.82	0.21
-0.125	-0.12	0.21	-0.82	0.23	-0.11	0.19	-0.71	0.21
0.125	0.13	0.21	-0.75	0.23	0.12	0.19	-0.60	0.21
0.385	0.36	0.21	-0.69	0.23	0.36	0.19	-0.51	0.22
0.670	0.63	0.21	-0.62	0.24	0.64	0.19	-0.41	0.22
1.040	1.00	0.21	-0.58	0.24	1.02	0.19	-0.36	0.22
1.640	1.56	0.20	-0.54	0.24	1.57	0.19	-0.31	0.23

Note: θ = Ability parameter used in generating normal item responses.

$\bar{\hat{\theta}}$ = Mean ability estimate.

$\overline{PSD}$ = Mean posterior standard deviation.

Discussion. Tables 1 and 2 document a potentially serious problem for short adaptive tests such as CAT-ASVAB. A sequence of as few as two random responses at the outset of the test may, in effect, anchor some estimates of ability far below true ability. The problem is not removed by administering additional items. Decisions based on these spuriously low ability estimates may lead to selection and classification errors with potentially serious implications for the examinee and the test user.

The negative consequences of a sequence of misinformative responses to initial CAT items warrant the development of methods for detecting the occurrence of this form of response aberrance. The posterior standard deviation (PSD), calculated during EAP estimation, would seem to be one logical candidate. A sequence of random responses would be expected to inflate the error in the ability estimate; the PSD should reflect this. From the results reported in Tables 1 and 2, however, it appears that the PSD was largely insensitive to the occurrence of random responses to initial CAT items.

III. STUDY 2: OPTIMAL DETECTION OF RANDOM RESPONSES TO INITIAL ITEMS

Purpose. The results in Tables 1 and 2 indicate that adaptive tests will underestimate the abilities of many examinees who give misinformative responses to initial items. This study examines the highest possible rates of detection for this form of response aberrance. Focusing on the upper bound of detectability allows for a determination of the potential for applying appropriateness measurement to this problem. Research addressing issues related to this application are justified only if reasonable detection rates have been observed in the optimal case.

The Appropriateness Index. Levine and Drasgow (1988) have shown that the Neyman-Pearson lemma (Lehman, 1959) can be used to obtain a most powerful statistic for testing the hypothesis that a response pattern is normal versus the hypothesis about a specific form of test-taking aberrance. For a vector of dichotomously scored responses u, the test statistic is

$$\lambda(\mathbf{u}) = P_A(\mathbf{u}) \,/\, P_N(\mathbf{u}), \tag{1}$$

where $P_A(\mathbf{u})$ is the probability of observing **u** when the response pattern was generated under conditions of aberrance (e.g., random responding to the first k items) and $P_N(\mathbf{u})$ is the probability of observing **u** when the response pattern was generated under normal test-taking conditions (e.g., the three-parameter logistic model).

For an n-item adaptive test with deterministic item selection

$$P_N(\mathbf{u}) = \int \Big(\prod_{i=1}^{n} P_{s_i}(\theta)^{u_i} \, [1 - P_{s_i}(\theta)]^{1-u_i} \Big) \, f(\theta) \, d\theta, \tag{2}$$

where s_i is the item number of ith item administered, $P_{s_i}(\theta)$ is the probability correct for an examinee with ability θ, u_i is the dichotomous response to the item administered at stage i, and $f(\theta)$ is the density of the ability distribution at θ.

To obtain $P_A(\mathbf{u})$, where aberrance in this case is random responding to the initial k items on the test, equation 2 is modified so that the random responses are modeled by $P_{s_i}(\theta) = .20$ (assuming 5-option items) for all θ and $i = 1$ to k:

$$P_A(\mathbf{u}) = \int \Big(\prod_{1}^{k} (.2)^{u_i} \, (.8)^{1-u_i} \Big) \Big(\prod_{k+1}^{n} P_{s_i}(\theta)^{u_i} \, [1 - P_{s_i}(\theta)]^{1-u_i} \Big) \, f(\theta) d\theta. \tag{3}$$

The appropriateness index based on equation 1, the likelihood ratio (LR), is the most powerful hypothesis test in the sense that maximum power is achieved for each given Type I error rate α. That is, in testing the hypothesis that a response pattern is "normal" versus the hypothesis that the pattern is "aberrant," no other appropriateness index computed from item responses provides greater power at α (Levine & Drasgow, 1988).

Appropriateness Index Power. One technique that has been used to examine the effectiveness of an appropriateness index in classifying normal and aberrant response patterns is the construction of Receiver Operating Characteristic (ROC) curves. Assuming that large index values are associated with aberrance, a point on the ROC curve is obtained by specifying a score t for the index and then computing

$x(t)$ = the proportion of population X (normal) response patterns with index values greater than t;

$y(t)$ = the proportion of population Y (aberrant) response patterns with index values greater than t.

The ROC curve consists of the points $(x(t), y(t))$ obtained for various values of t. The false alarm rate (probability of incorrectly labeling a normal pattern as aberrant) when using t as the cut score is given by $x(t)$. The hit rate (probability of correctly identifying an aberrant pattern) is given by $y(t)$.

Asymptotic ROC Curves. Previous examinations of the detection rates of various appropriateness indices used distributions of normal and aberrant response patterns generated with Monte Carlo procedures, combining these patterns into a single group and ordering them on the magnitude of the appropriateness index, and constructing an ROC curve in the manner described in the preceding paragraph. These sample ROC curves represent stochastic approximations to the population ROC. The accuracy of these approximations has not been extensively investigated.

An analytic procedure for constructing ROC curves is presented in the following paragraph. This procedure determines the asymptotic ROC curve for a specified appropriateness index, form of aberrance, and adaptive test. Unlike the sample-based ROC curves generated by Monte Carlo methods, the asymptotic ROC curves contain no sampling error and thus display the performance of the appropriateness index in the population.

Specifying a test length of n items and an item pool for the deterministic adaptive test, the analytic procedure involves the following steps. First, all possible 2^n dichotomous response patterns are enumerated. For each response pattern $\mathbf{u} = (u_1, u_2, \ldots, u_n)$, the n items that the adaptive test would administer to an examinee demonstrating pattern $\mathbf{u}$ are then determined. The a, b, and c parameters for these items and $\mathbf{u}$ are then used to compute $P_N(\mathbf{u})$ and $P_A(\mathbf{u})$. An exact value for k is specified for $P_A(\mathbf{u})$. At this point in the procedure, an appropriateness index is computed for response pattern $\mathbf{u}$. In the case of the LR index, the probabilities $P_N(\mathbf{u})$ and $P_A(\mathbf{u})$ (equations 2 and 3) are used to obtain λ.

Since a 15-item deterministic adaptive test has only $2^{15} = 32,768$ different response patterns, it is possible to compute $P_N(\mathbf{u})$ and $P_A(\mathbf{u})$ exactly for each pattern $\mathbf{u}$. This property allows for constructing the asymptotic ROC curve. By ordering the response patterns from highest to lowest on the magnitude of the appropriateness index (assuming that large index values are associated with response aberrance) and computing cumulative P_N and P_A with each successive pattern, the information required to construct the asymptotic ROC curve is obtained.

For $\mathbf{u}^1$, the response pattern associated with the largest index value, $P_A(\mathbf{u}^1)$ and $P_N(\mathbf{u}^1)$ represent the values $y[\lambda(\mathbf{u}^1)]$ and $x[\lambda(\mathbf{u}^1)]$, respectively. These coordinates give the hit rate and false alarm rate, respectively, and represent the initial point on the ROC curve. The second point on the ROC curve represents the hit and false alarm rates when using cut score $\lambda(\mathbf{u}^2)$, where $\mathbf{u}^2$ is the response pattern associated with the second largest index value. A point on the ROC curve is obtained by computing $x[\lambda(\mathbf{u}^2)] = P_N(\mathbf{u}^1) + P_N(\mathbf{u}^2)$ and $y[\lambda(\mathbf{u}^2)] = P_A(\mathbf{u}^1) + P_A(\mathbf{u}^2)$. The coordinates for the kth point are $x[\lambda(\mathbf{u}^k)] = P_N(\mathbf{u}^1) + P_N(\mathbf{u}^2) + \ldots + P_N(\mathbf{u}^{k-1}) + P_N(\mathbf{u}^k)$ and $y[\lambda(\mathbf{u}^k)] = P_A(\mathbf{u}^1) + P_A(\mathbf{u}^2) + \ldots + P_A(\mathbf{u}^{k-1}) + P_A(\mathbf{u}^k)$, where $\mathbf{u}^k$ is the response pattern giving the kth largest value of λ. In this way, the 2^n points for the asymptotic ROC curve are obtained.

Analyses. A series of analyses of the type described in the preceding paragraphs were performed to examine the power of the LR index for detecting random responses to the first k items of a 15-item adaptive test, where k equaled 1, 2, 3, 4, or 5. All possible response patterns for a 15-item test were enumerated, resulting in 32,768 patterns. The item pool for the test

consisted of the 100 most informative items from the 258 items in the CAT-ASVAB Word Knowledge item pool. Modal Bayesian estimation (Owen, 1969, 1975) was used to obtain ability estimates at each stage of the test and items were selected to maximize information at $\hat{\theta}$. Equations 2 and 3 were computed using a normal [0,1] density and Simpson's rule for numerical integration. Asymptotic ROC curves were then constructed for the LR index in each of the five aberrance conditions.

Results. Table 3 contains selected points from the asymptotic ROC curves for the LR index. Figure 1 shows the ROC curves. The detection rates given in Table 3 indicate that an initial sequence of random responses can be accurately identified. High hit rates, relative to low false alarm rates, were obtained using the LR index. High detection rates were achieved even in the least extreme case of aberrance. When aberrance is defined as a random response to the first item only, the LR index correctly identifies over half of the true instances of aberrance at a false alarm rate of .10. Approximately the same degree of power exists for the extreme case, where random responses are given for the first five items, but this level of detectability is achieved at the expense of a false alarm rate of only .001.

Discussion. Table 3 and Figure 1 present a successful application of appropriateness measurement to a form of response aberrance that can have serious consequences for CAT-ASVAB scores. The results in Table 3 indicate that an initial sequence of random responses can be detected with high levels of accuracy. These detection rates were achieved at the false alarm levels that would be required for practical application.

The results of this study demonstrate the potential for appropriateness measurement in CAT. These results justify further studies focusing on practical issues involved in implementing an appropriateness measurement procedure for detecting misinformative responses to initial CAT items.

Table 3.

Proportion of Aberrant Response Patterns Detected by the Likelihood Ratio Index at Selected ROC Curve Points (Aberrance = Random Responding to the Initial k Items)

α	k = 1	k = 2	k = 3	k = 4	k = 5
.001	.13	.26	.35	.45	.49
.005	.21	.36	.46	.56	.62
.01	.27	.41	.52	.61	.66
.02	.33	.47	.57	.66	.71
.03	.37	.50	.61	.69	.75
.04	.40	.53	.64	.71	.77
.05	.43	.55	.66	.73	.78
.07	.47	.59	.69	.76	.81
.10	.52	.64	.73	.80	.85

Note: α = Proportion of Normal Patterns Incorrectly Identified as Aberrant.

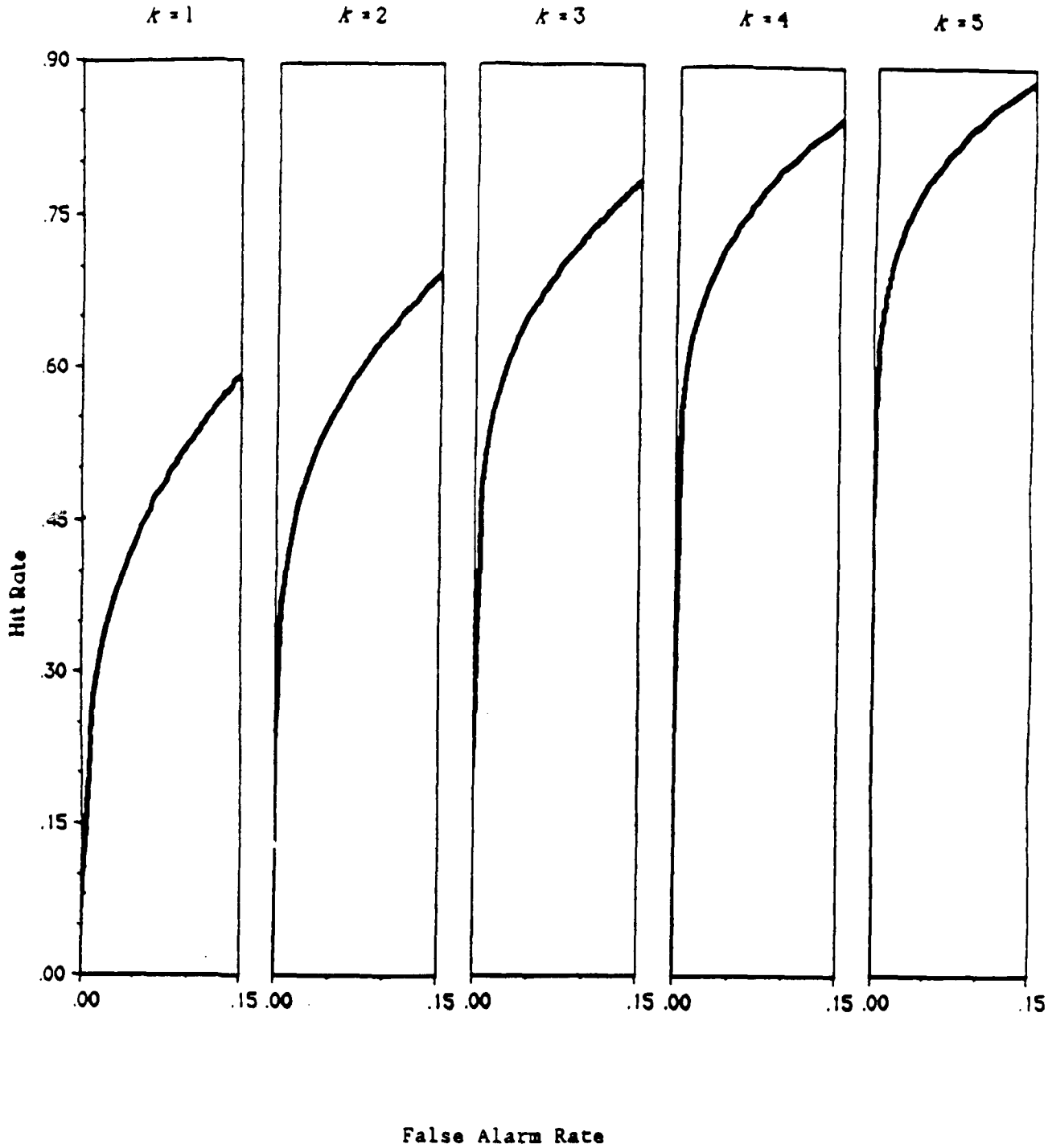

Figure 1. ROC Curves for the Likelihood Ratio Index Where Aberrance = Random Responses to the Initial *k* Items.

IV. STUDY 3: DETECTION RATES OF NONOPTIMAL INDICES

Purpose. The effectiveness of several different appropriateness indices will be evaluated in this study. The indices include Drasgow, Levine, and Williams' (1985) standardized version of the log likelihood index (Levine & Rubin, 1979), Tatsuoka's (1984) "extended" caution index, and two fit statistics given by Rudner (1983). These four indices were chosen because they have been successful in detecting some forms of response aberrance when applied to conventional tests. In some cases, these indices have provided detection rates approaching the optimal levels given by the LR index (Drasgow et al., 1987). Each of these nonoptimal indices is relatively easy to compute, thereby increasing their appeal as candidates for practical testing situations. The primary goal of this study is to assess the degree to which these indices are less than optimal.

Appropriateness Indices. Drasgow et al. (1985) provided an approximate standardization of Levine and Rubin's (1979) L_0 index. They computed

$$L_0 = \Sigma [u_i \ln P_i(\hat{\theta}) + (1-u_i) \ln Q_i(\hat{\theta})], \quad (4)$$

where $\hat{\theta}$ is an ability estimate computed from the examinee's item responses, $P_i(\hat{\theta})$ is the probability of an examinee with ability $\hat{\theta}$ making the correct response to item i, u_i is the dichotomous item response for item i, and $Q_i(\hat{\theta}) = 1-P_i(\hat{\theta})$. Drasgow et al. (1985) give an approximate standardization of L_0 as:

$$LZ = \frac{L_0 - E(L_0)}{[Var(L_0)]^{1/2}} \quad (5)$$

where

$$E(L_0) = \Sigma [P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + Q_i(\hat{\theta}) \ln Q_i(\hat{\theta})], \quad (6)$$

and

$$Var(L_0) = \Sigma [P_i(\hat{\theta})Q_i(\hat{\theta})][\ln(P_i(\hat{\theta})/Q_i(\hat{\theta}))]^2. \quad (7)$$

In previous studies with conventional paper-and-pencil tests, maximum likelihood ability estimates (MLEs) were used to provide the $\hat{\theta}$s of equations 4 through 7. In the present study of adaptive computer administered tests, modal Bayesian estimates are used

Rudner (1983) suggested two indices that are three-parameter generalizations of Wright's (1977) Rasch model fit statistics:

$$F1 = \frac{1}{n} \sum \frac{[u_i - P_i(\hat{\theta})]^2}{P_i(\theta)Q_i(\theta)} \tag{8}$$

and

$$F2 = \frac{\Sigma [u_i - P_i(\hat{\theta})^2]}{\Sigma P_i(\theta)Q_i(\theta)}. \tag{9}$$

The last appropriateness index is the approximate standardization of the fourth "extended caution index" given by Tatsuoka (1984):

$$T4 = \frac{\Sigma [P_i(\hat{\theta}) - u_i)(P_i(\hat{\theta}) - P)]}{\Sigma P_i(\hat{\theta})Q_i(\hat{\theta})(P_i(\theta) - P)^2]^{1/2}} \tag{10}$$

where

$$P = \frac{1}{n} \Sigma P_i(\hat{\theta}). \tag{11}$$

Analyses. Asymptotic ROC curves for LZ, F1, F2, and T4 were constructed using the analytic procedure described in Study 2. Test specifications were identical to those used in Study 2. Aberrance was defined as random responses to the initial five items of the test.

Results. Table 4 gives the proportion of aberrant patterns detected by each index at various false alarm rates between .00 and .10. Also listed in Table 4 are the optimal detection rates given by the LR index. Figure 2 shows the ROC curves for each of the nonoptimal indices.

From the results in Table 4 it is evident that none of the nonoptimal indices provided aberrance detection approaching optimal levels. At an error rate of .005, for example, the best performing nonoptimal index (F2) was only 25% as powerful as the LR index.

Discussion. The nonoptimal indices represent omnibus tests of response aberrance. In this sense, their ineffectiveness in detecting random responses to the intial five items of a 15-item test is not surprising. The LR index provides superior detection because an alternative hypothesis is specified.

Table 4 Proportion of Aberrant Response Patterns Detected by 5 Appropriateness Indices at Selected ROC Curve Points (Aberrance = Random Responses to the Initial 5 Items)

α	LR	LZ	F1	F2	T4
.001	.49	.00	.01	.11	.01
.005	.62	.11	.03	.16	.02
.01	.66	.12	.05	.20	.09
.02	.71	.15	.08	.22	.13
.03	.75	.18	.11	.23	.14
.04	.77	.20	.19	.25	.16
.05	.78	.23	.20	.26	.20
.07	.81	.26	.23	.28	.24
.10	.85	.29	.27	.32	.27

Note: α = Proportion of Normal Patterns misclassified as Aberrant.
LR = Likelihood Ratio.
LZ = Standardized Log Likelihood.
F1 = Standardized Squared Residual.
F2 = Standardized Squared Residual.
T4 = Standardized Extended Caution Index.

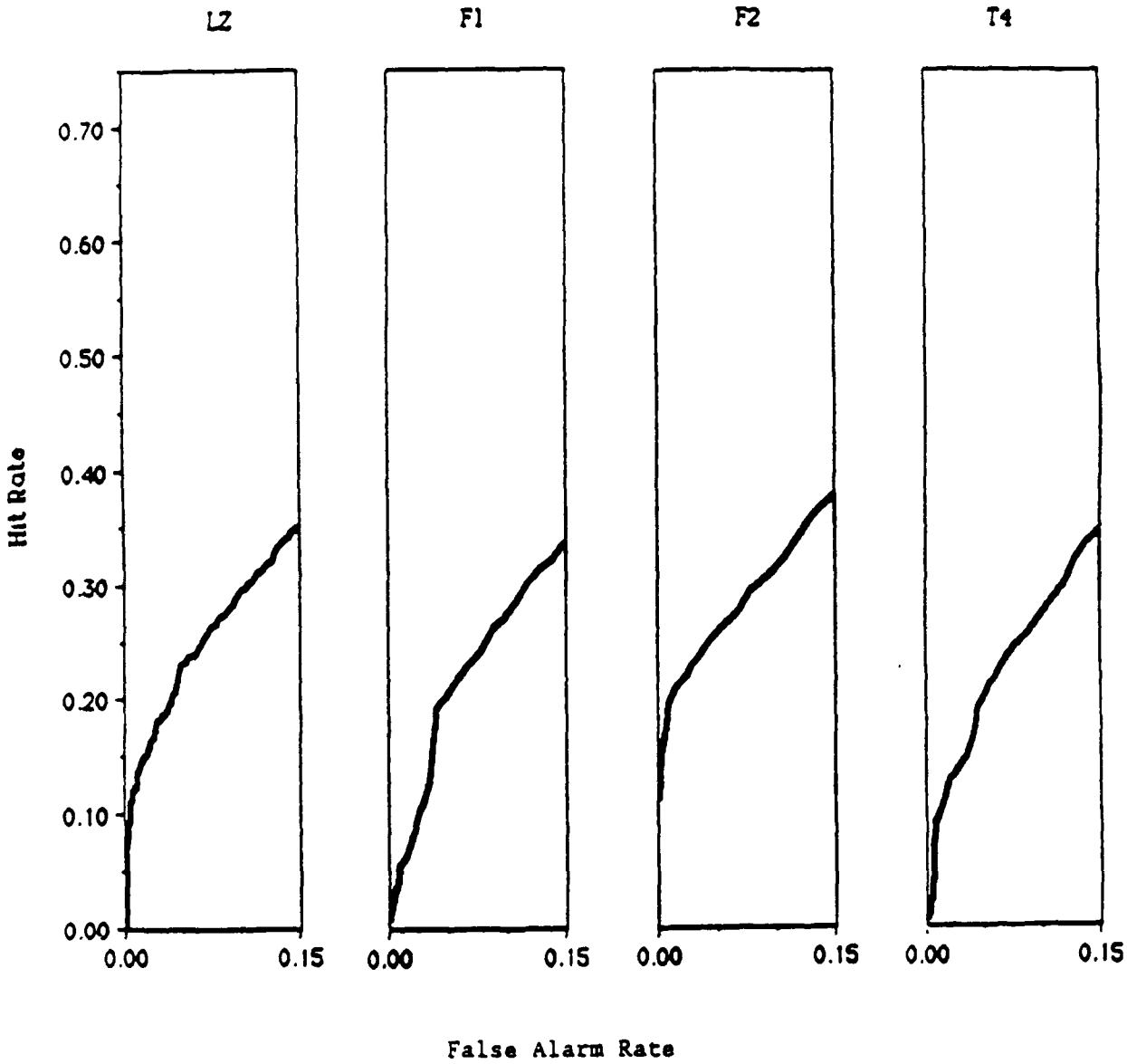

Figure 2. ROC Curves for Nonoptimal Appropriateness Indices.

Another constraint to aberrance detection was provided by the small number of item responses analyzed. Comparisons of results for LZ, F1, F2, and T4 on an 85-item test (Drasgow et al., 1987) and a 30-item test (Drasgow, Levine, Williams, McLaughlin, & Candell, in press) show a significant decrease in detecting the 30% spuriously low condition with the shorter test. In applying these indices to 15-item tests, further decreases in detection rates would be expected. Molenaar and Hoijtink's (1987) warning against the use of nonoptimal indices on tests of fewer than 20 items appears to be valid in the present case. Five aberrant responses and 10 nonaberrant responses do not appear sufficient to generate a signal distribution that is distinguishable from the distribution of noise when using an omnibus index such as LZ or T4.

V. STUDY 4: RECOVERY OF ASYMPTOTIC ROC CURVES USING MONTE CARLO METHODS

Purpose. The ROC curves presented in Studies 2 and 3 were constructed by an analytic, as opposed to probabilistic, procedure and contain no sampling error; these curves give the performance of the appropriateness indices in the population. In contrast, previous appropriateness measurement studies have used Monte Carlo methods to generate samples of normal and aberrant response patterns prior to constructing ROC curves. Two important concerns with these methods are the accuracy and stability of the ROC curves they produce. To date, the sampling behavior of ROC curves constructed by Monte Carlo methods has not been systematically examined. The sample sizes for normal and aberrant response patterns needed for accurate and stable ROC curves are unknown.

This study focuses on the accuracy and variability of sample-based ROC curves. Monte Carlo methods of the type found in previous appropriateness measurement studies are used to generate samples of aberrant and normal patterns. ROC curves are constructed using different Ns for aberrant and normal samples to examine the effects of sample size.

The results of these analyses provide important information for at least two reasons. First, previous research on appropriateness measurement can be reexamined to see if the sample sizes used in these studies provide accurate results in the present analyses. Determining the Ns needed for accurate and stable ROC curves will also make it possible to interpret with greater confidence the Monte Carlo studies presented in Study 5 and Study 7. Although the analytic procedure is an elegant and powerful approach for constructing ROC curves under certain limiting conditions (e.g., tests of 15 items or less), several questions about detection rates cannot be addressed with this approach but instead require Monte Carlo simulation. One of these questions involves the effect of an item security algorithm on detection rates, which is the focus of Study 5. Study 5 involves a larger population of response pattern/test items combinations than can be handled with the analytical procedure, given practical constraints such as computer memory and computer processing time.

Data Generation. Normal and aberrant response patterns were generated by first sampling an ability parameter from a normal distribution with mean

0.0 and variance of 1.0. At each of the 15 stages of the adaptive test, the probability of making a correct response was given by the three-parameter logistic function. Aberrant response patterns were created by fixing the probability of a correct response at 0.20 for the first five items. The item pool for the test consisted of 100 CAT-ASVAB Word Knowledge items. Ability estimates were obtained using modal Bayesian estimation, and items were selected at each stage of the test by using a maximum information criterion. The initial ability estimate was set to 0.0, the mean of the Bayesian prior, for each simulated examinee.

Samples of normal patterns consisted of 200, 1,000, 2,000, or 4,000 patterns. Aberrant samples were 100, 500, 1,000, or 2,000. These sample sizes were chosen to span the range used in previous research. Table 5 provides a summary of the 16 conditions.

Analyses. ROC curves were constructed by first merging the samples of normal and aberrant patterns and ordering these patterns from highest to lowest with respect to the LR index computed for the pattern. Beginning in the list of patterns with the aberrant pattern associated with the largest index value, and proceeding with each successive aberrant pattern, the proportion of aberrant patterns and normal patterns existing at that level or above in the list were determined, providing the coordinates for a point on the ROC curve. Thus, the hit rate was uniformly incremented by the reciprocal of the aberrant sample N at each point on the ROC curve.

The procedure for constructing ROC curves described in the preceding paragraph differs from the procedure implicit in the definition of an ROC curve given in Study 2. As defined in Study 2, ROC curves are constructed by evaluating hit rates and false alarm rates at each index value, ordered from largest to smallest (assuming that large index values are associated with aberrance). Using this approach, each successive point on the ROC curve (i.e., moving from low false alarm rates to high false alarm rates) is not necessarily associated with a larger y-coordinate. For example, if all of the patterns for two successive index values are nonaberrant, the hit rate will remain constant while the false alarm rate increases. The contrasting approach used in the present study, where each successive point on the ROC curve necessarily *does* possess a larger y-coordinate than the previous point, was chosen for reasons of computational efficiency. It is

Table 5. Summary of Data Sets used to Evaluate the Recovery of Exact ROC Curves

Data Set	Aberrant Sample Size	Normal Sample Size
1	100	200
2	500	200
3	1,000	200
4	2,000	200
5	100	1,000
6	500	1,000
7	1,000	1,000
8	2,000	1,000
9	100	2,000
10	500	2,000
11	1,000	2,000
12	2,000	2,000
13	100	4,000
14	500	4,000
15	1,000	4,000
16	2,000	4,000

important to note that the fundamental shape of the ROC curve will be the same using either procedure described in Study 2 or the procedure used in the present study.

Conditions 1-3, 5-7, and 9-11 contained 20 independent replications. To reduce CPU time, conditions 4, 8, and 12-16 contained 10 independent replications. A graphic procedure was used to examine the variability of the Monte Carlo ROC curves. This procedure is similar in spirit to Thissen and Wainer's (1983) "N-line plot" technique for evaluating the confidence envelopes for item characteristic curves. For each condition, the ROC curves generated by Monte Carlo simulation were plotted simultaneously. Recovery of the asymptotic ROC curve was also evaluated graphically by plotting points from this curve against the sample-based curves.

Results. Figure 3 shows the ROC curves constructed using 200 normal patterns and 100, 500, 1,000, and 2,000 aberrant patterns, respectively. Figure 4 shows the curves for samples of 1,000 normal patterns and each of the four sample sizes for aberrant patterns. Figures 5 and 6 give the curves constructed using 2,000 and 4,000 normal patterns, respectively, and the four sample sizes for aberrant patterns. The open circles in each plot are points from the asymptotic ROC curve for the LR index.

It is evident from Figures 3 through 6 that the range of sample sizes used in the 16 conditions resulted in large differences in the variability and accuracy of the ROC curves. It is also clear that the smaller sample sizes do not provide acceptable levels of stability and accuracy. Conditions 1, 5, 9, and 13 (100 aberrant patterns) resulted in the least accurate ROC curves. The most dramatic improvement in accuracy and stability occurs when the aberrant sample size is increased from 100 to 500 (conditions 2, 6, 10, and 14), but the ROC curves generated in these conditions are still highly variable.

Condition 16 (4,000 normal, 2,000 aberrant) provided the most accurate and stable ROC curves. Eight of the 10 curves in this condition very closely approximated the asymptotic ROC curve. Condition 12 (2,000 normals, 2,000 aberrants) and Condition 15 (4,000 normals, 1,000 aberrants) appear to be equivalent with respect to accuracy, but the curves constructed in condition 15 are somewhat less variable.

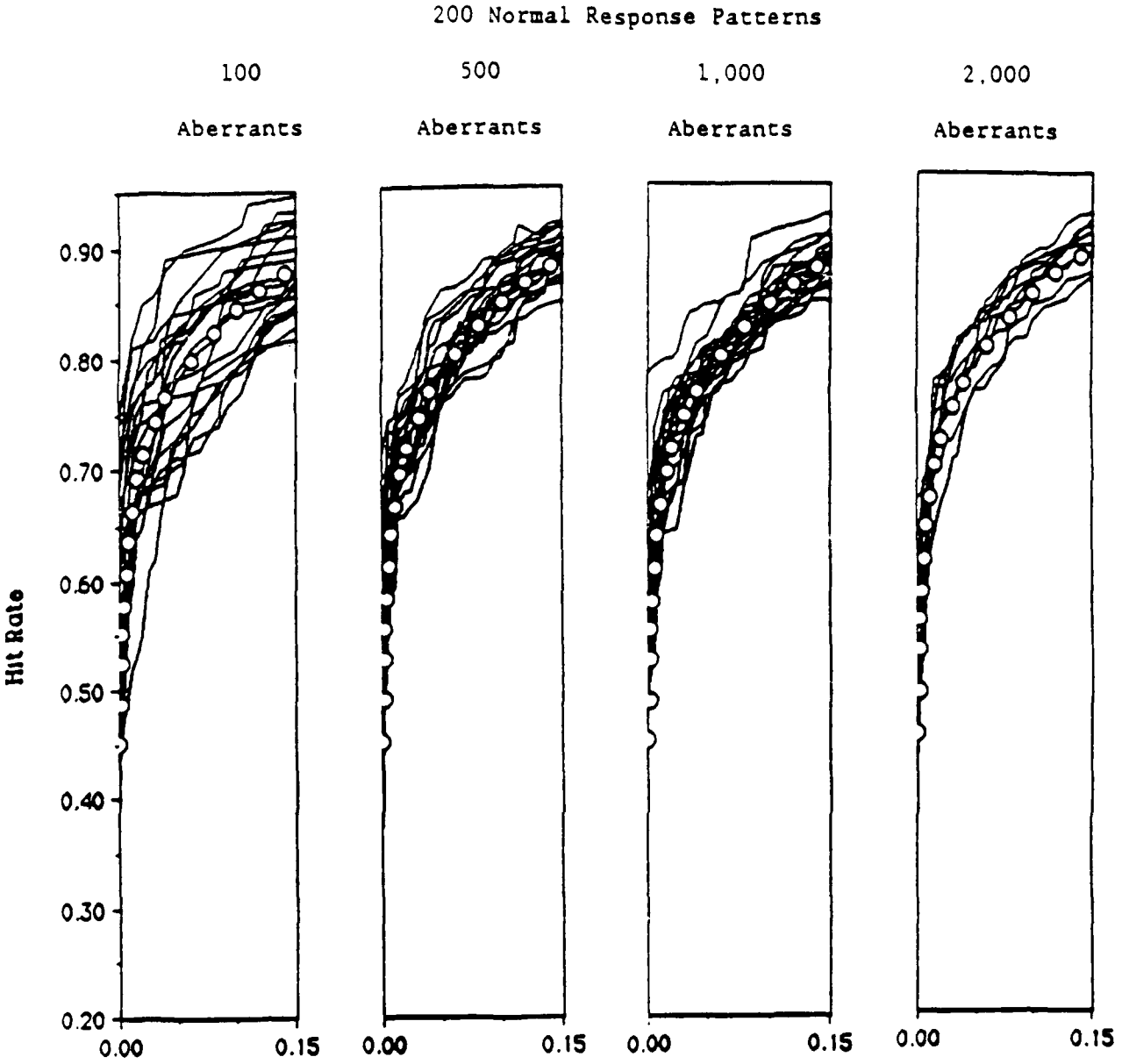

Figure 3. ROC Curves for Data Sets 1 - 4.

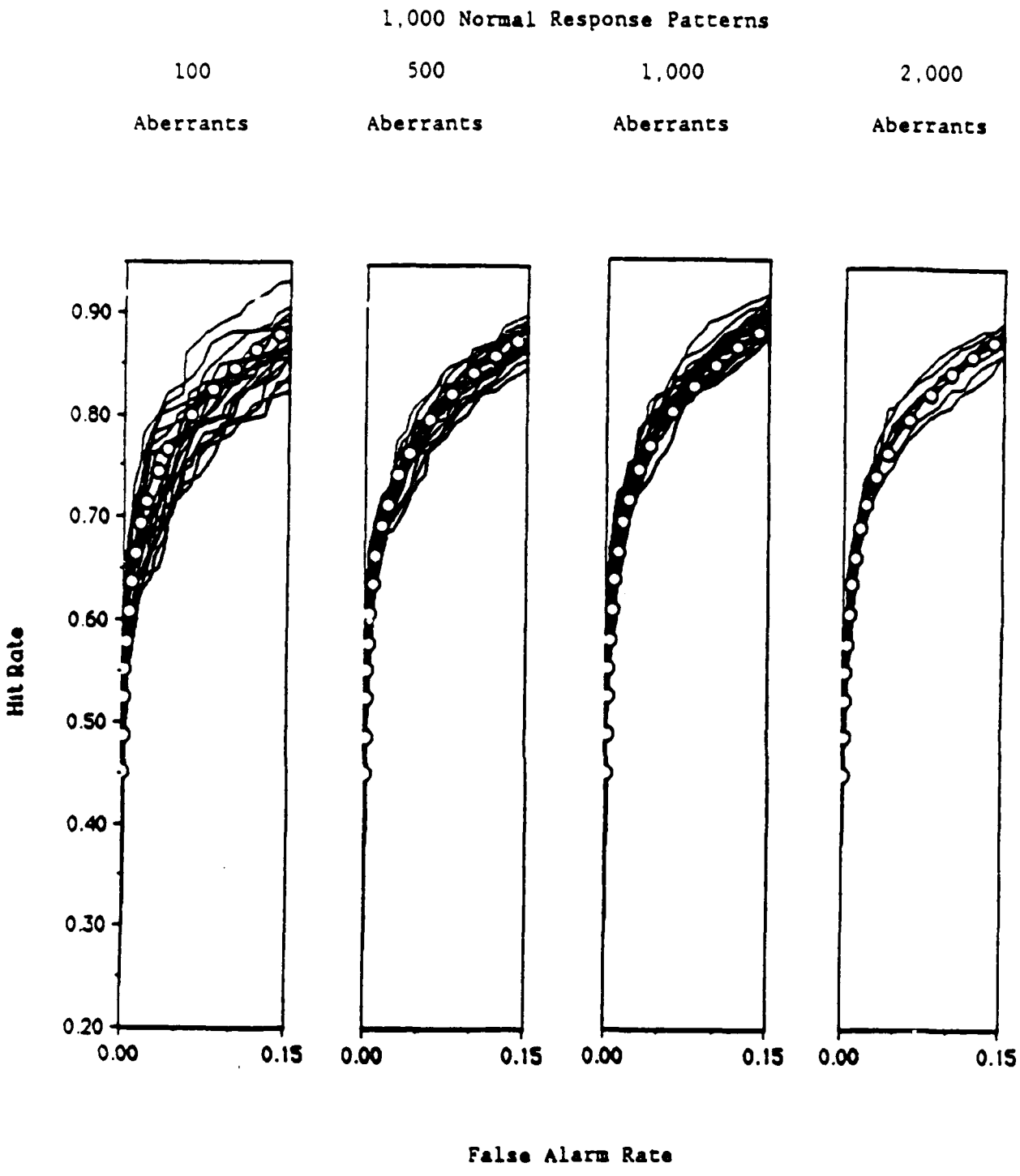

Figure 4. ROC Curves for Data Sets 5 - 8.

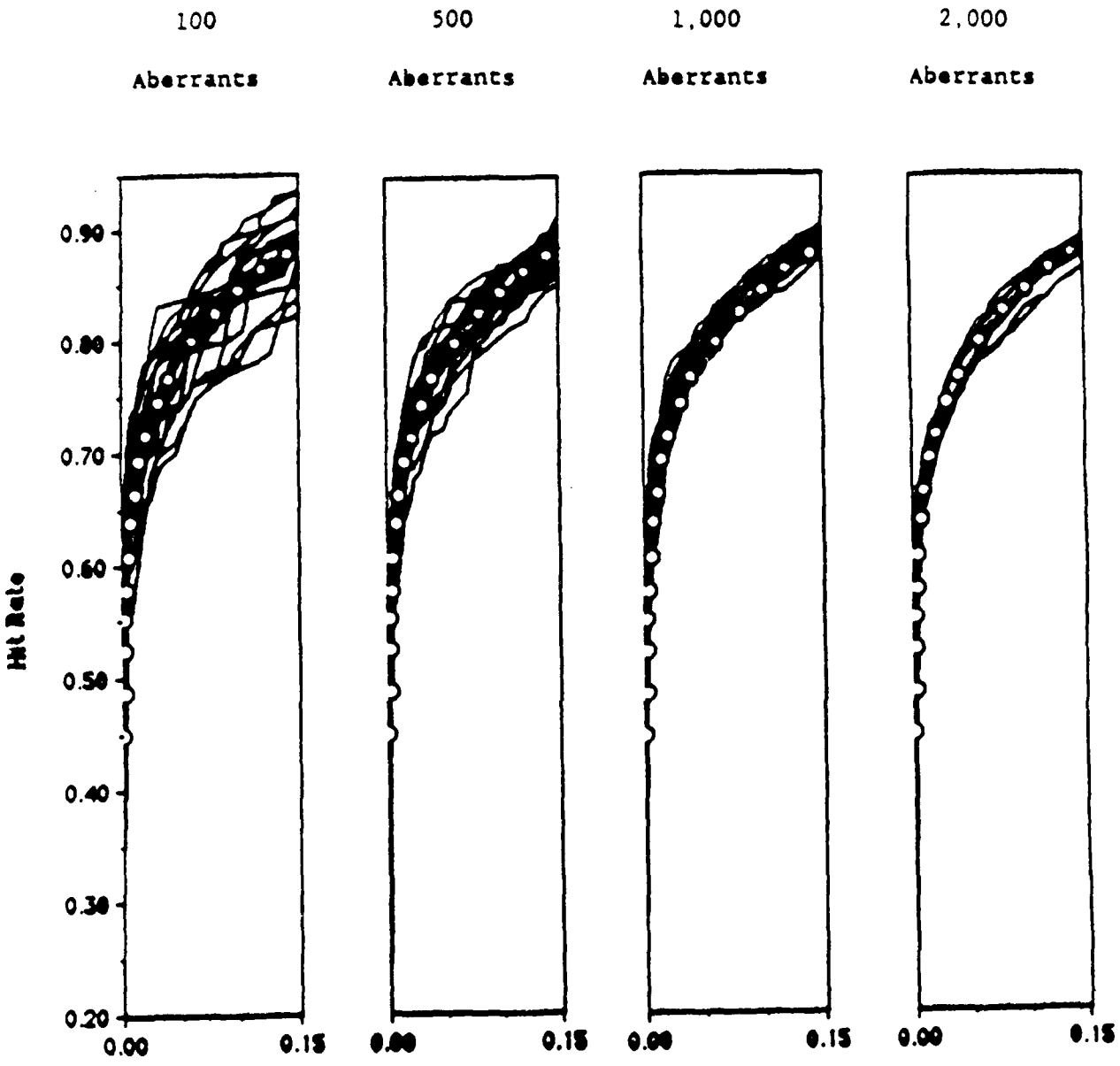

Figure 5. ROC Curves for Data Sets 9 - 12.

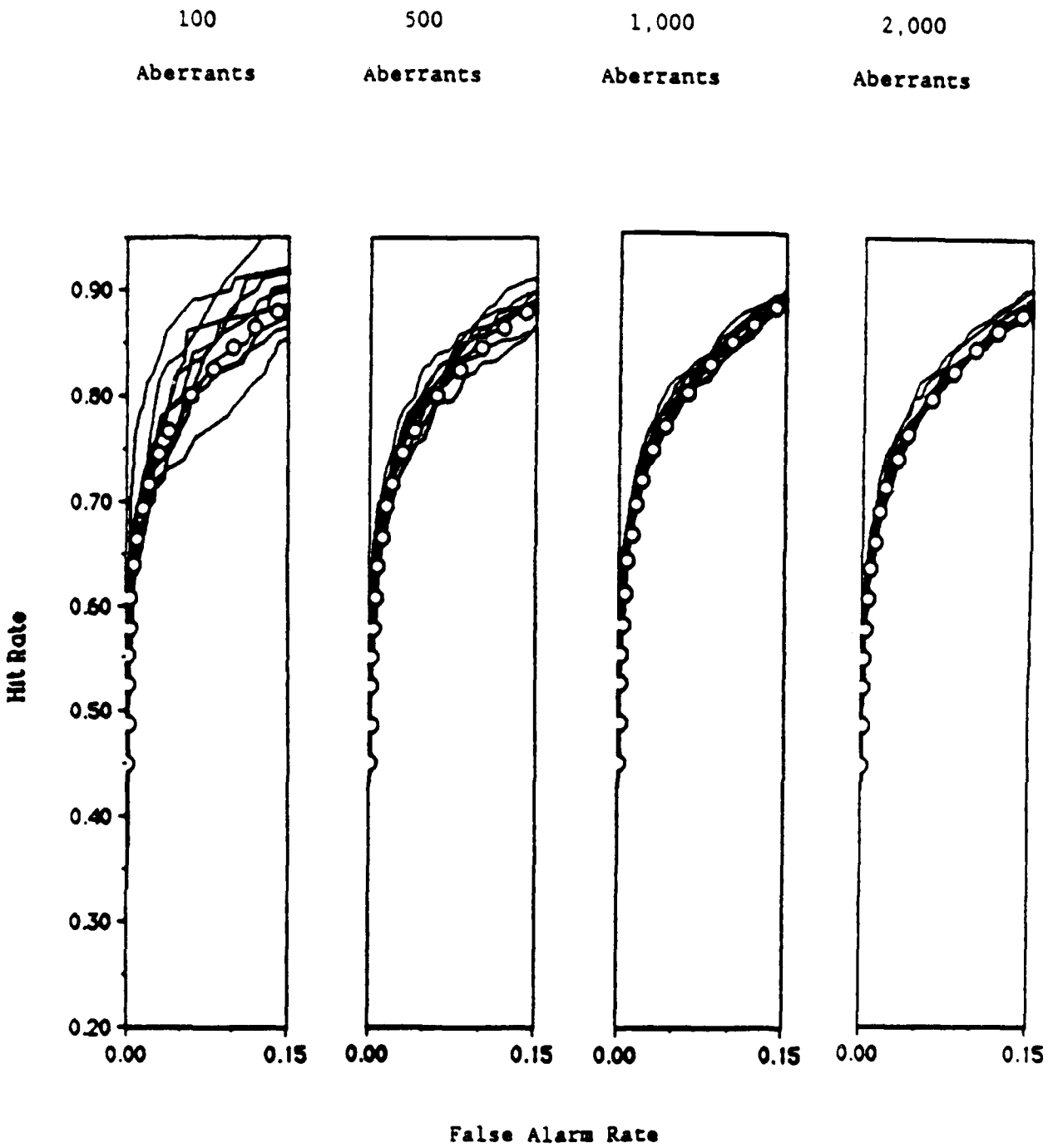

Figure 6. ROC Curves for Data Sets 13 - 16.

Discussion. The purpose of the analyses presented in this study was to determine the necessary sample sizes for constructing ROC curves using the LR index. Samples of 4,000 normal patterns and 2,000 aberrant patterns provide accurate approximations to the asymptotic ROC curve. These sample sizes will be used in Studies 5 and 7.

Samples of 4,000 normal and 2,000 aberrant patterns have been used frequently in previous appropriateness measurement research. The results presented in this study provide at least partial support for the Monte Carlo methods used in these earlier studies. It is important to note that the present results were obtained for a test of 15 items. The interaction of test length and sample size needs to be examined with longer tests. However, constructing asymptotic ROC curves using the number of items found in many standardized tests is not currently feasible; computer memory and processing requirements for such analyses are prohibitive. For example, there are more than 35 million dichotomous response patterns for a 25-item test; over 1 billion patterns exist for a 30-item test. An alternative to the analytic procedure will be required for evaluating the asymptotic ROC curves for longer tests.

VI. STUDY 5: EFFECT OF AN ITEM SECURITY PROCEDURE ON OPTIMAL DETECTION

Purpose. The results presented in Study 2 for the LR index are based on an adaptive test that is deterministic with respect to item administration at each branch of the test. That is, a one-to-one correspondence exists between each dichotomous response pattern and the set of items administered by the test. It is not likely that this situation would be encountered in practice, however. For purposes of test security, many adaptive tests use procedures for limiting the exposure of items.

The question addressed in this study is whether an item security procedure degrades the power of the LR index for detecting random responses to initial items. If aberrance detection changes significantly as a result of the item security algorithm, the analytic procedure used in Studies 2 and 3 cannot be used to evaluate the LR index with practical adaptive tests. Instead, Monte Carlo procedures of the type used in Study 4 will be required. If the performance of LR is unaffected by the item security algorithm, the analytic approach will remain a valid procedure for evaluating the index with adaptive tests.

Data Generation. Monte Carlo procedures identical to those described in Study 4 were used to create samples of 4,000 normal and 2,000 aberrant response patterns. Aberrant patterns contained random responses to the first five items of the test; response probabilities for the remaining 10 items were determined by the three-parameter logistic model. The item pool and ability estimation procedures for the adaptive test were identical to those used in Studies 1 through 4. Unlike the test used in Studies 1 through 4, however, the adaptive test simulated in the present study used an item security procedure when selecting items. At each stage of the test, the most informative unused item is sampled with a probability of .25 from the appropriate row in the information table. If this item is not selected, the next most informative unused item is sampled with probability .33. If neither of the first two items is selected, the third most informative item is sampled with probability .50. The fourth most informative item is administered if none of the first three items are sampled.

Ten independent sets of aberrant and normal samples were generated. ROC curves were constructed within each replication in the manner described

in Study 4, using the LR index to identify aberrance. These curves were then plotted simultaneously, along with points from the asymptotic ROC curve for the LR index.

Results. Figure 7 shows the 10 ROC curves for the LR index where a security procedure was used when selecting items. The open circles in Figure 7 are points from the asymptotic ROC curve for LR, where the most informative unused item was selected at each stage of the test.

The ROC curves in Figure 7 display more variability than the curves shown in condition 16 of Figure 6, which were constructed with identical sample sizes but without an item security procedure. Taken together, the curves in Figure 7 indicate that the item security procedure did not reduce the power of the LR index. Five of the curves show higher detection rates than the asymptotic ROC curve for the LR index. The other five curves closely approximate the asymptotic ROC curve.

Discussion. Item security procedures are a necessary feature for large-scale adaptive CAT. These procedures minimize the exposure of items in the item pool, thereby decreasing the likelihood that examinees may benefit from previous experience with the same test. In addition, new items will not need to be developed as often when the exposure of current items is limited by a security procedure.

Results from the present study indicate that the item security procedure used in selecting CAT-ASVAB Word Knowledge items does not decrease the power of the LR index when compared to detection rates obtained without the security procedure. Of course it is possible that other security procedures may greatly decrease detectability, but in view of the similarity between various proposed security procedures, this seems unlikely. Consequently it seems safe to conclude that LR will detect about as well with a deterministic administration procedure as with a security procedure.

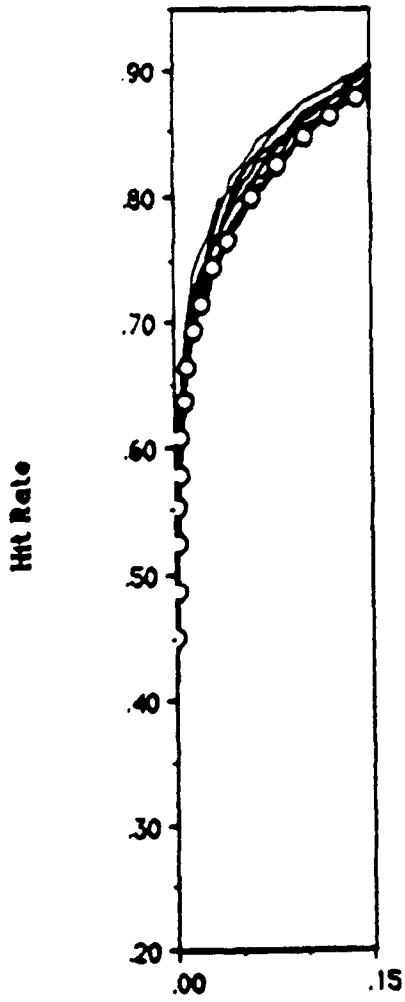

Figure 7. ROC Curves Generated Using an Item Security Algorithm.

VII. STUDY 6: STANDARDIZATION OF THE LIKELIHOOD RATIO INDEX

Purpose. Observed differences between distributions of appropriateness index scores for normal and aberrant patterns may reflect differences in ability or test score distributions in addition to providing evidence of index effectiveness. An index value at one ability level or test score may suggest a high probability of response aberrance while the same index value at another test score indicates an adequate fit of the IRT model. In such a case, the index confounds appropriateness measurement with ability.

The problem identified in the previous paragraph involves the standardization of an appropriateness index, which has been defined as the extent to which the conditional distributions of the index differ across the range of θ for non-aberrant examinees (Drasgow et al., 1987). Well-standardized indices display equivalent conditional distributions across ability levels. These indices allow for a single cutting score and provide aberrance detection that is largely independent of differences in ability distributions across normal and aberrant samples. Poorly standardized indices make it impossible to use a single cutting score to classify response patterns, since a given index value can have dramatically different meaning depending on the corresponding test score or ability. Standardized interpretations of these index scores require evaluations of the conditional distributions of the index.

The LR index is not well standardized. To date no attention has been given to standardizing LR because the index has been used as a benchmark statistic and not as a practical appropriateness index. Since any attempt to standardize LR will lower detection rates, the LR index has been used in its unstandardized form in order to provide optimal detection rates. This study examines detection rates for a standardized version of LR.

Analyses. A standardized version of the LR index was developed by evaluating cumulative distributions of the logarithm of LR within several mutually exclusive ability intervals. Asymptotic distributions of log LR were obtained by enumerating all possible response patterns, in the same manner used to construct asymptotic ROC curves in Studies 2 and 3. The probability of an observed log LR score, conditioning on estimated ability, served as the standardized index. This procedure will be described in detail in the following paragraphs.

There are two approaches to studying the standardization of an appropriateness index. The scientific question of the relation between the LR index and ability can be addressed by focusing on changes in $F(\lambda|\theta)$ across θ, whereas a more practical approach focuses on changes in $F(\lambda|\hat{\theta})$ across $\hat{\theta}$. It seems important to investigate the practical approach because conditioning will take place on estimated ability in practice. Therefore, conditional distributions of LR were defined on intervals of $\hat{\theta}$.

Asymptotic conditional distributions were created by first enumerating all possible patterns for a 15-item test, computing LR for each pattern, and ordering these patterns on $\hat{\theta}$. An important decision that has to be made at this point in the analysis is how to divide the complete distribution of 32,768 index scores into conditional distributions. One of the goals in the present study was to accurately document the changes in the distribution of LR across $\hat{\theta}$. Ideally, the number of $\hat{\theta}$ intervals and span of each interval should be determined in such a way that the observed changes in the conditional distributions are gradual. For comparison purposes, these distributions should contain equivalent numbers of patterns. Conditional distributions of LR were created in the present study by selecting successive sets of 1,024 patterns from the ordered list. In this way, 32 distributions of equal size were created, with each distribution containing approximately 3% of the response patterns.

The logarithm of the LR index (hereafter referred to as LLR) was computed for each pattern to reduce the variance of index scores across conditional distributions. Note that this is a monotone transformation; the rank orderings of LR are preserved in the transformation.

A standardized version of the LLR index was created by transforming each index score into an approximately uniformly distributed score. This was accomplished by computing the LLR score and $\hat{\theta}$ for each response pattern and then calculating the conditional probability of the observed LLR score given that the ability estimate falls into $\hat{\theta}$'s interval. Thus, the standardized LLR index is:

$$LLR_s = \text{Prob}\ (LLR \leq LLR' \mid \hat{\theta} \in I), \qquad (12)$$

where LLR' is the observed index score for a normal response pattern and I is the interval containing the $\hat{\theta}$ for the pattern. If for each ability interval I the conditional distribution $F_I(t) = \text{Prob}(LLR \leq t \mid \hat{\theta} \in I)$ is well

approximated by a continuous, strictly increasing distribution function, then LLR_s and the random variable giving the $\hat{\theta}$ interval of the response pattern are approximately independent.

The effectiveness of the standardized LLR index in detecting random responses to the initial five items was evaluated by constructing an asymptotic ROC curve. A comparison of this ROC curve with the asymptotic ROC curve for the LR index provides evidence of the impact of standardization on the power of LR.

Results. Table 6 shows the LLR values at several cumulative probabilities in each of the 32 conditional distributions used in the study. Evaluation of these distributions was restricted to the left tail for practical reasons; large index values will be used as cutting scores. In several cases, no empirical value for LLR existed at a specific probability and linear interpolation or extrapolation was required.

LLR (and thus LR) is not well standardized. The left tails of the conditional distributions of LLR vary considerably across $\hat{\theta}$ intervals. At the same time, the 32 distributions do a reasonable job of documenting the changes in LLR as a function of $\hat{\theta}$.

From Table 6 it is apparent that a large percentage of the response patterns do not yield LR scores greater than 1.00. The response patterns with the highest 3% of ability estimates, for example, will produce LR values greater than 1.00 fewer than 5 times in 1,000. Response patterns giving $\hat{\theta}$s in the range [0.95, 1.26] will *never* produce LR indices greater than 1.00. In general, $\hat{\theta}$s greater than 0.57 are extremely unlikely when the initial five responses to the test are random. This result indicates that patterns yielding these $\hat{\theta}$s can be excluded from appropriateness analyses without significantly reducing power.

Response patterns giving $\hat{\theta}$s in the range [-0.81, -0.64] and [-1.21, -1.14] are much more likely to produce large index scores. Patterns within these ranges of $\hat{\theta}$ are more likely to be classified as aberrant than are patterns giving other $\hat{\theta}$s. Consequently, the detection rates at these $\hat{\theta}$ levels will be accompanied by higher false alarm rates. Table 7 gives the hit rates and false alarm rates for the unstandardized LR index and the standardized LLR index. At low false alarm rates, the power of the LR index is greatly reduced by standardization. The standardized index classifies

Table 6. LLR Index Scores at Various Cumulative Proportions.

$\hat{\theta}$ Interval	Alpha Level .001	.005	.01	.05	.10
[1.26, + ∞]	1.05	-1.85	-2.56	-4.17	-4.52
[0.95, 1.26]	-0.28	-0.49	-1.03	-1.95	-3.23
[0.80, 0.95]	2.25	1.05	-1.70	-1.98	-2.11
[0.68, 0.80]	4.94*	0.66	0.48	-1.94	-2.06
[0.57, 0.68]	0.77*	-0.21	-0.27	-1.28	-1.96
[0.42, 0.57]	3.16	1.09	-0.49	-1.01	-1.31
[0.18, 0.42]	3.60*	0.76	-0.24	-1.15	-1.28
[-0.02, 0.18]	4.46	2.08	1.32	-0.67	-1.10
[-0.15, -0.02]	4.61*	2.58	1.77	0.18	-0.81
[-0.24, -0.15]	6.51	4.01	2.27	0.36	-0.31
[-0.32, -0.24]	4.61	2.36	2.13	0.88	0.43
[-0.40, -0.32]	5.38*	4.01	2.03	0.51	0.20
[-0.47, -0.40]	3.39	2.95	1.36	0.42	0.01
[-0.56, -0.47]	6.71*	5.95	2.60	0.26	-0.06
[-0.64, -0.56]	2.10*	2.02	1.98	1.17	-0.30
[-0.72, -0.64]	7.40*	3.75	3.29	1.30	0.01

Note: * = extrapolated value (no empirical value with conditional probability < alpha).

Table 6 (Continued)

$\hat{\theta}$ Interval	Alpha Level .001	.005	.01	.05	.10
[-0.81, -0.72]	7.20*	5.23	2.91	2.12	1.05
[-0.90, -0.81]	4.48	3.93	2.35	1.58	0.96
[-0.98, -0.90]	4.30*	3.96	3.53	1.27	1.10
[-1.06, -0.98]	2.91*	2.88*	2.84	1.77	0.88
[-1.14, -1.06]	2.17*	2.17*	2.15	1.49	0.63
[-1.21, -1.14]	11.97*	6.00	1.88	1.43	1.13
[-1.29, -1.21]	1.98	1.39	1.25	0.89	0.56
[-1.38, -1.29]	5.37*	5.28	5.10	0.67	0.49
[-1.48, -1.38]	5.02*	4.52	3.61	0.46	0.29
[-1.59, -1.48]	3.34*	3.21	3.20	2.40	0.43
[-1.71, -1.59]	2.56*	2.54	2.52	2.28	1.94
[-1.83, -1.71]	1.76*	1.75	1.72	1.62	1.49
[-1.92, -1.83]	1.99	1.14	1.12	1.05	0.97
[-2.05, -1.92]	1.09	1.06	0.83	0.69	0.62
[-2.22, -2.05]	1.08	0.88	0.84	0.40	0.38
[- ∞, -2.22]	1.12	0.90	0.51	0.38	0.26

Note: * = extrapolated value (no empirical value with conditional probability < alpha).

Table 7. Proportion of Aberrant Response Patterns Detected by Standardized and Unstandardized Indices at Selected ROC Curve Points

α	Standardized LLR Index	Unstandardized LR Index
.001	.00	.49
.005	.16	.62
.01	.52	.66
.02	.60	.71
.03	.68	.75
.04	.72	.77
.05	.75	.78
.07	.80	.81
.10	.85	.85

Note: α = Proportion of Normal Patterns Misclassified as Aberrant.

less than 1% of the aberrant patterns correctly at a false alarm rate of .001, whereas the unstandardized index has a 50% detection rate at the same error level. At a false alarm rate of .005, the standardized index has only 25% of the power of the unstandardized index. The detection rates for both indices are roughly equivalent at false alarm rates of .05 and beyond.

Discussion. The analyses presented in this study represent only one of several approaches to standardizing appropriateness indices. The intent here was not to develop the best standardization of the LR index but rather, to evaluate the effect of one type of empirical standardization on the power of LR. Molenaar and Hoijtink (1987) and Kogut (1988) provide examples of alternative procedures for obtaining the conditional distributions of an appropriateness index.

The need for standardization results from a confounding of test scores or θ and appropriateness measurement. Several forms of response aberrance will produce a strong relation between estimated ability and measured appropriateness. For example, we would expect that when cheating takes place over a significant portion of a test, the resulting test scores or ability estimates will most often be above average. Consequently, when attempting to identify cases of cheating, non-aberrant response patterns that yield above-average scores are more likely to be misclassified as aberrant than are non-aberrant patterns yielding below-average scores. Because labeling an examinee as a possible cheater carries potentially serious consequences for that examinee, detection procedures that are more likely to misclassify one group of examinees than other groups will not be considered fair. Use of a well-standardized appropriateness index, however, will create a situation where no single group of examinees--classified by ability--is "singled out" when testing for cheating. Although the standardized index may provide less power than its unstandardized counterpart, the decrease in detection rates may be a necessary cost for ensuring that measured appropriateness and test scores are independent.

In other circumstances, the unstandardized index may be preferred over the standardized version. This might be the case when detection rates decrease significantly after standardizing and the consequences of using the unstandardized index--in which higher false alarm rates will occur for certain ability groups--are not considered to be particularly negative for

examinees or test users. In general, procedures that identify some examinees as possibly having test scores that are spuriously low, with the concomitant opportunity to obtain a higher test score, should be well received by most of these examinees.

The negative consequences of random responses to initial items on an adaptive test result from the test's inability to recover from this situation. In these cases, the examinee has given a pattern of responses from which the adaptive test cannot adequately estimate ability. It may be somewhat of a misnomer to label these response patterns as aberrant; what is certain is that these patterns will often produce spuriously low test scores. Identifying examinees who give these patterns focuses attention on the test itself rather than on the examinees. The social stigma that results from being suspected of other forms of response aberrance, notably cheating, should not result in this case. Thus, the cost of higher false alarm rates for some ability groups that will result when using the unstandardized LR index may be an acceptable price to pay in order to obtain optimal detection rates.

VIII. STUDY 7: GENERALIZABILITY OF DETECTION USING THE LIKELIHOOD RATIO INDEX

Purpose. Likelihood ratio tests achieve their power through specificity. An optimal test for detecting four initial random responses is not an optimal test for detecting five initial random responses. The purpose of this study is to determine how sensitive LR is to misspecification of the number of initial random responses. In the unlikely event that a likelihood ratio test for four initial random responses turns out to be nearly optimal for detecting any number of initial random responses, then a very simple, nearly optimal appropriateness measurement procedure can be formulated.

On the other hand, if it turns out that the optimal test for k initial responses is not close to being optimal in detecting $k' \neq k$ initial random responses, then the situation is somewhat more complicated. Nonetheless an optimal test can still be obtained by first determining the proportion of people who respond randomly to the first item only, to the first two items, to the first three items, etc. The point is discussed after some results are presented.

Study 7 attempts to determine the ability of an LR index designed to detect random responding on the first k items to detect random responding on a shorter or longer initial string of items.

Data Generation. Four conditions of actual aberrance were simulated by creating samples of 2,000 response patterns where the initial two, three, four, or five responses were random. These aberrant samples were each combined with 4,000 normal patterns and LR indices for detecting random responses to the initial two, three, four, or five items were computed for each pattern. ROC curves were then constructed for each of the 16 conditions defined by the combinations of actual and hypothesized aberrance.

Results. Figure 8 shows three graphs containing the ROC curves for the LR index computed to detect random responses to the initial two items. Actual aberrance consisted of random responses to the initial three, four, and five items. The darker curve in each graph is the ROC curve for the truly optimal index. Optimal ROC curves represent the detection rates that would be achieved if the actual form of aberrance were correctly specified by the LR index. Figure 9 compares the ROC curves for the LR index for

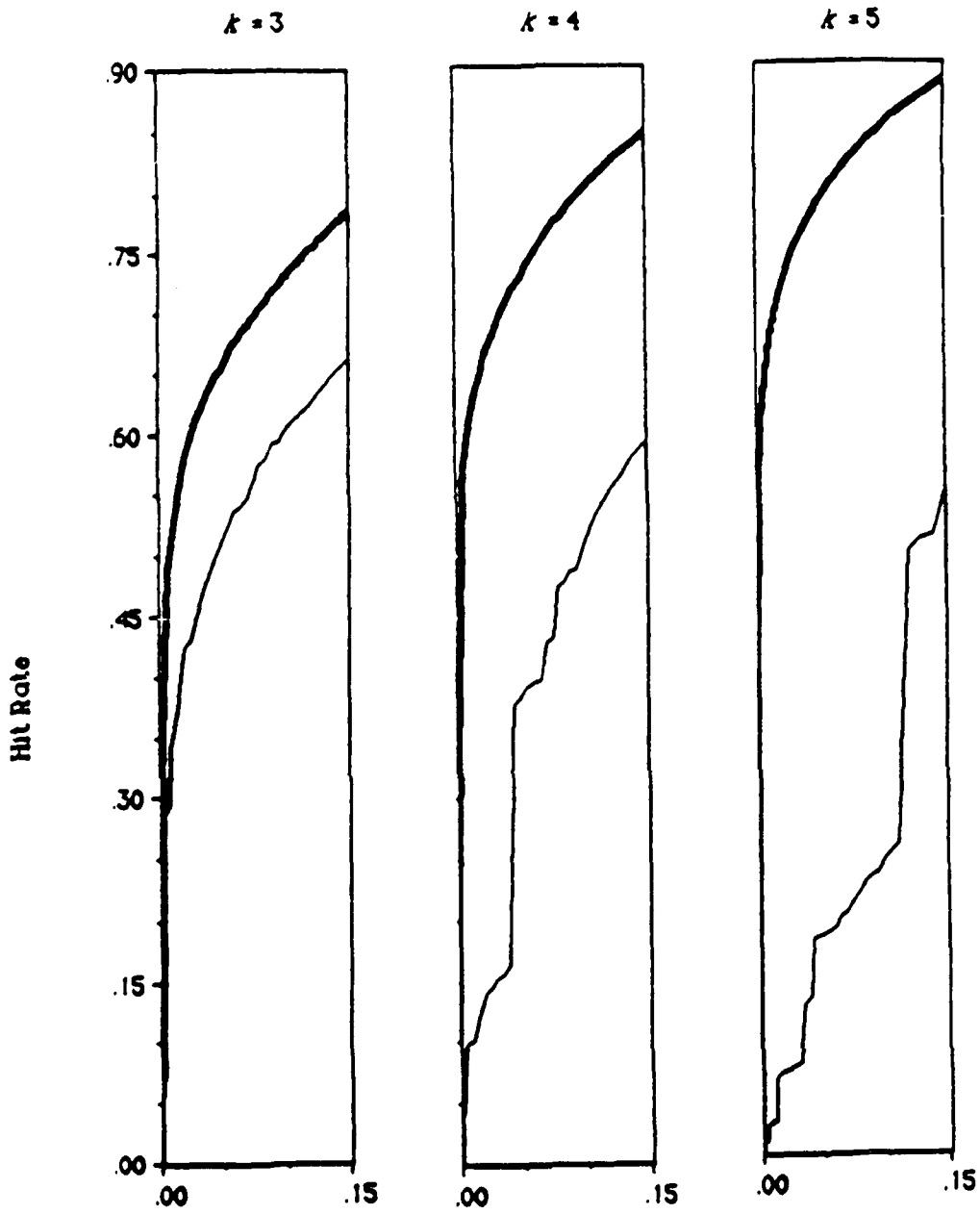

Figure 8. ROC Curves for Likelihood Ratio Index where Hypothesized Aberrance = Random Responses to the Initial Two Items and Actual Aberrance = Random Responses to the Initial k Items.

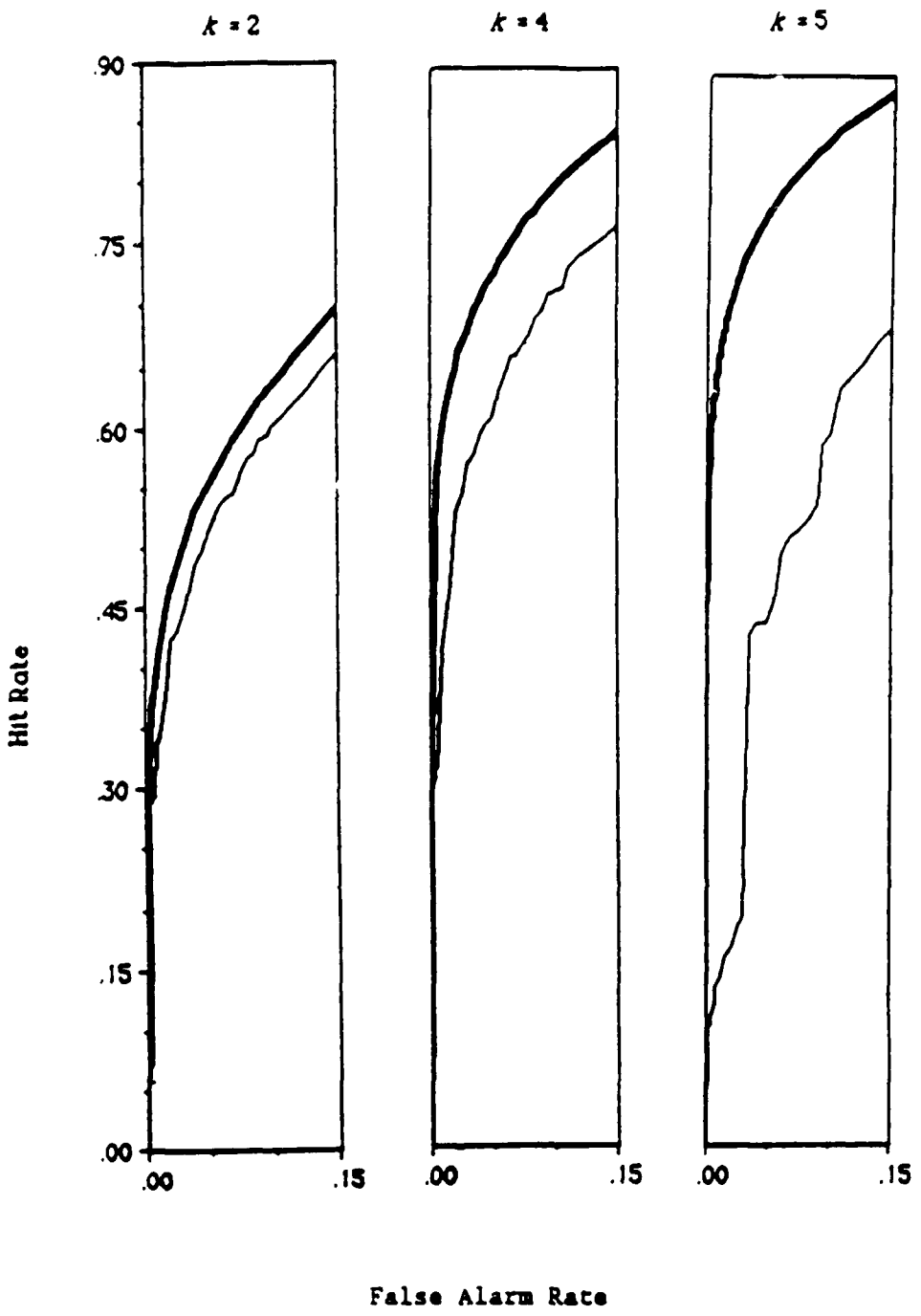

Figure 9. ROC Curves for Likelihood Ratio Index where Hypothesized Aberrance = Random Responses to the Initial Three Items and Actual Aberrance = Random Responses to the Initial k Items.

detecting three random responses with optimal detection rates, where actual aberrance was two, four, and five random responses. Figure 10 shows the ROC curves for the LR index for detecting four random responses, where actual aberrance was two, three, and five random responses, versus optimal detection rates. Figure 11 shows the ROC curves using the LR index for five random responses, where aberrance was actually random responses to the initial two, three, and four items, compared to optimal ROC curves.

The LR index computed for two random responses (Figure 8) is reasonably effective at detecting three random responses, but is much less effective at detecting initial sequences of four and five random responses. The LR index computed for three random responses (Figure 9) performs well when actual aberrance is either two or four random responses. The LR index for four random responses (Figure 10) is effective in detecting either three or five random responses. The LR index for five random responses (Figure 11) does a good job at detecting random responses to the initial four items.

Discussion. The high detection rates displayed by the LR indices appear to generalize to adjacent forms of aberrance. Together the LR indices for three or four items give moderately high detection rates for initial random responses on two through five items. That is, the LR index computed for three random responses (Figure 9) does a good job of identifying patterns containing random responses to the initial two or four items and provides optimal detection for the case of three random responses, while LR computed for four responses (Figure 11) detects an initial sequence of three, four, or five random responses well.

The likelihood ratio indices show sufficient sensitivity to length of the initial segment of random responses to be useful for estimating the proportion of people fumbling on one item, on two items, on three items, etc. When these proportions are known, a single optimal test can be formulated as follows. Let LR_k denote the likelihood ratio statistic for random responding on exactly k items. Let p_k denote the proportion of people randomly responding on exactly k items. Thus p_0 is the proportion of normal examinees, p_1 the proportion of examinees randomly responding on the first item only, etc. Then it can be shown that the index

$$LR_p = \sum_{k=1}^{15} p_k \, LR_k \qquad (13)$$

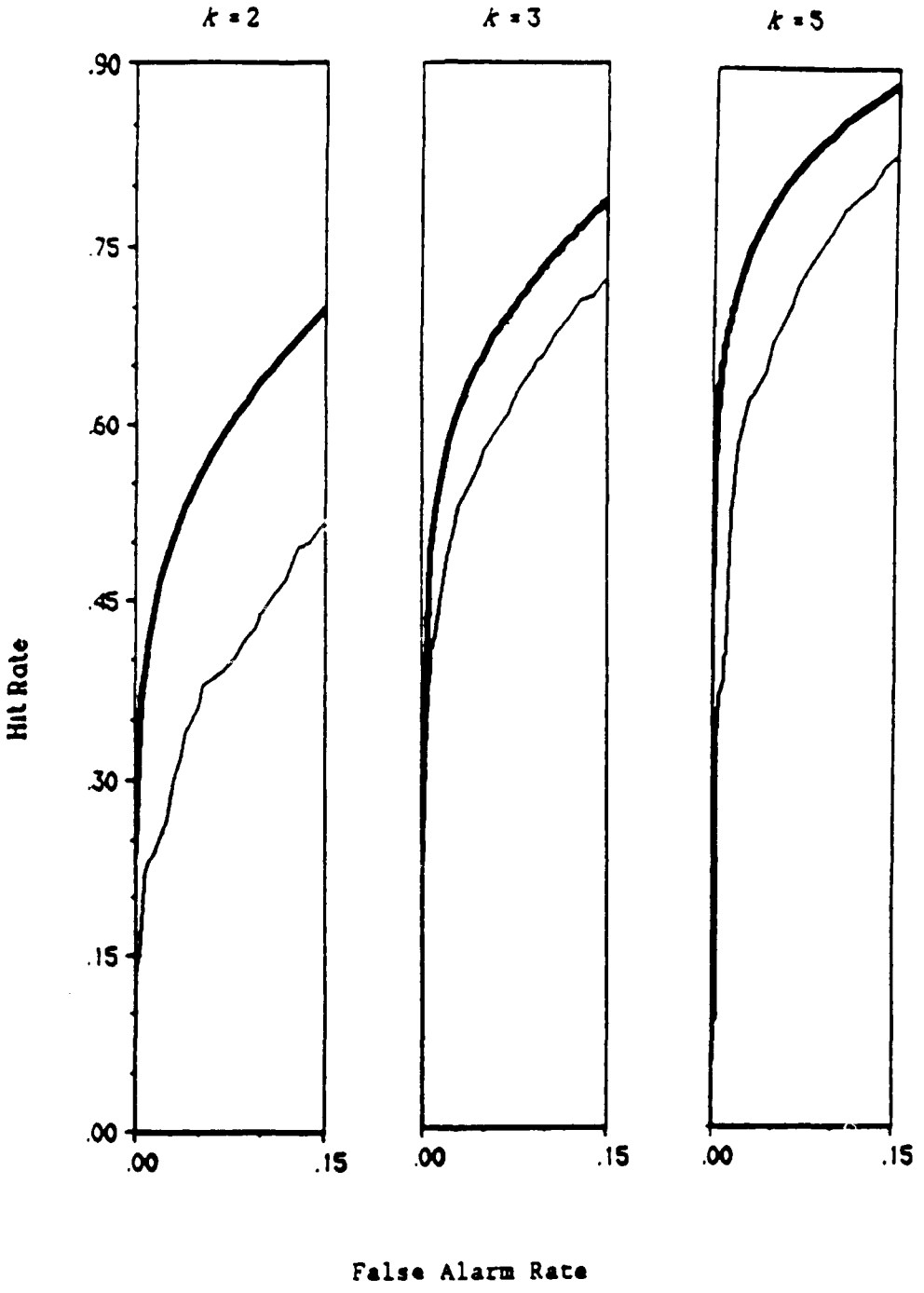

Figure 10. ROC Curves for Likelihood Ratio Index where Hypothesized Aberrance = Random Responses to the Initial Four Items and Actual Aberrance = Random Responses to the Initial k Items.

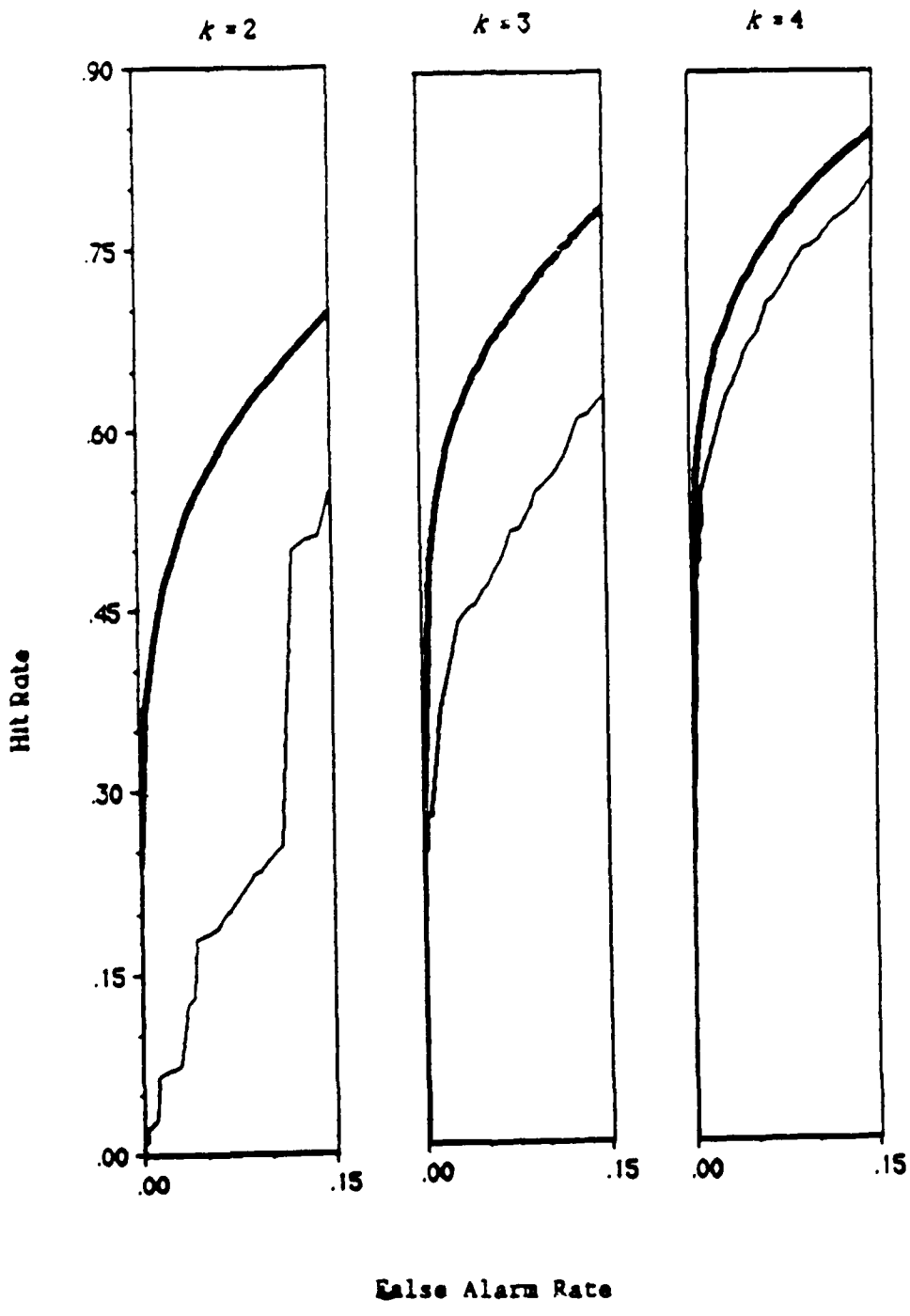

Figure 11. ROC Curves for Likelihood Ratio Index where Hypothesized Aberrance = Random Responses to the Initial Five Items and Actual Aberrance = Random Responses to the Initial k Items.

is optimal for testing the hypothesis that the examinee responded randomly on at least one item against a hypothesis of no fumbling. This index is optimal in the sense that it produces the highest ROC curve, or equivalently, has highest power among all tests for fumbling with any specified false positive rate.

IX. CONCLUSIONS

CAT has the potential for providing accurate ability estimation with significantly fewer items than standardized tests, provided the IRT model used in estimating ability and selecting items is appropriate. The responses used to compute initial estimates of ability are particularly important. When these responses fail to conform to the IRT model, final ability estimates are found to be highly inaccurate, even when subsequent items are answered in accordance with the IRT model. The results presented in Study 1 indicate that random responses to as few as the initial two items establish an ability estimate that anchors the final $\hat{\theta}$ well below θ. For examinees of above-average ability, the amount of underestimation can be severe.

The LR index, which provides the theoretically highest possible detection rates for a specified degree of aberrance, was shown in Study 2 to have significant power for detecting an initial sequence of random responses. High hit rates, relative to low false alarm rates, were observed for even the least severe case where only the initial response was random. These results indicate that a sequence of random responses to initial items on an adaptive test can be detected with the degree of accuracy required by test practitioners.

Four nonoptimal appropriateness indices were examined in Study 3. Unlike the LR index, these indices do not provide a specific test of an alternative hypothesis, but several have demonstrated near optimal detection rates for some forms of aberrance on conventional tests. These nonoptimal indices were not effective, however, in detecting random responses to the initial five items on a 15-item adaptive test.

In Study 4, sample-based ROC curves were constructed varying Ns for normal and aberrant patterns. These analyses were needed to interpret the accuracy of previous studies using sample-based ROC curves and to establish the sample sizes needed for Study 5 and Study 7. The results from Study 4 indicated that samples of 4,000 normal patterns and 2,000 aberrant patterns provided for near-asymptotic ROC curves.

Study 5 examined the power of the LR index when an item security procedure was used during item selection. The item security procedure used in Study 5 did not reduce the effectiveness of the LR index. Consequently,

the security procedure was not used in subsequent studies. Note that the computation of LR is not complicated by the introduction of a security procedure. The analytic procedure for estimating ROC curves is not used with a nondeterministic item selection algorithm because there are too many possible test outcomes.

The LR index was shown to be poorly standardized in Study 6. A standardized version of LR was developed, with the standardized index providing significantly less power than the unstandardized index at low false alarm rates. Fortunately, standardization is not nearly so important in the detection of fumbling as it is with socially undesirable behavior, such as cheating.

Finally, the results in Study 7 indicate that the LR index is capable of detecting sequences of random responses adjacent to the sequence specified when computing the index. In particular, the LR index computed for random responses to the initial three items was able to identify a sequence of two or four random responses adequately, the index computed for four items was able to detect either three or five reasonably well, and LR computed for five items displayed high levels of power for detecting a sequence of four random responses. The sensitivity of LR to the length of the initial segment of random responses suggested LR may be useful in determining the distribution of aberrance.

Recommendations for Further Research

1. Obtaining Base Rates for Aberrance. The LR indices possess remarkable power for identifying patterns containing sequences of random responses of specified length. Thus, the indices may be useful in estimating the base rate or distribution of fumbling in the general population and in various groups. With such estimates the effectiveness of procedures designed to reduce fumbling could be evaluated. Estimates could also be used to determine whether fumbling is more common in minority or other demographically defined groups. Finally, as noted in Study 7, the relative frequency of different degrees of fumbling can be incorporated in a single optimal test for any degree of fumbling.

2. Decision-Theoretic Appropriateness Measurement. Drasgow and Guertler (1987) have argued that the detection rates for an appropriateness index in

simulation studies provide only a portion of the information needed to determine a cut score for the index. The selection of a cut score should also incorporate the base rate of aberrance in the population of interest as well as the relative importance of the possible classification outcomes. For example, which outcome has greater disutility--classifying a normal pattern as aberrant (false positive) or classifying an aberrant pattern as normal (false negative)? The values assigned to these potential outcomes will likely have a significant effect on how the appropriateness index is evaluated and used.

Drasgow and Guertler (1987) suggest a Bayesian scheme for handling the problem of selecting a cut score, where an index score t is classified as aberrant when

$$\frac{\text{Prob}(t|\text{Ab})}{\text{Prob}(t|\text{N})} > \frac{1 - \text{Prob}(\text{Ab})}{\text{Prob}(\text{Ab})} \times \frac{U_4 - U_2}{U_1 - U_3} \qquad (14)$$

where Prob(t|Ab) / Prob(t|N) is the likelihood ratio (equation 1), Prob(Ab) is the base rate of aberrance, and U1,U2,U3, and U4 are the (dis)utilities associated with the possible classification outcomes:

U1 classifying an aberrant pattern as aberrant ("correct positive");

U2 classifying a normal pattern as aberrant ("false positive");

U3 classifying an aberrant pattern as normal ("false negative");

U4 classifying a normal pattern as normal ("correct negative").

Comparisons of index scores developed using equation 14 in different testing situations may highlight interesting and important distinctions in how appropriateness measurement might be implemented.

3. Evaluating Change Scores. An initial sequence of random responses to an adaptive test using Bayesian estimation produces a $\hat{\theta}$ that typically underestimates θ for examinees of average to above-average ability. For high-ability examinees, the underestimation will be severe. For less able examinees, the negative bias in the ability estimate will be less severe. In practice, we would like to identify those examinees whose θs suffer the most from random responses to initial items.

In simulation research we can determine the change score for an examinee by computing

$$\hat{\theta}_C = \hat{\theta}_N - \hat{\theta}_A, \quad (15)$$

where $\hat{\theta}_N$ is the ability estimate obtained when each item in the adaptive test is answered in accordance with the IRT model and $\hat{\theta}_A$ is the ability estimate obtained after the examinee gives random responses to initial items. With this additional information, the distribution of $\hat{\theta}_C$ among the aberrant patterns *not* identified by LR can be observed. It is expected that $\hat{\theta}_C$ will generally be small among aberrant examinees not identified by LR. We have observed that response patterns that are hard to detect typically have small change scores.

REFERENCES

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

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 4, 431-444.

Broadbent, D.W. (1971). Decision and stress. New York: Academic Press.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.

Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. Journal of Applied Psychology, 72, 10-18.

Drasgow, F., & Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. Applied Psychological Measurement, 10, 59-67.

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. Applied Psychological Measurement, 11, 59-79.

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-80.

Drasgow, F., Levine, M.V., Williams, B., McLaughlin, M.E., & Candell, G.L. (In Press). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. Applied Psychological Measurement.

Easterbrook, J.A. (1959). The effect of emotion on cue utilization and the organization of behavior. Psychological Review, 66, 183-201.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory: Applications to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Hunt, E., & Pellegrino, J. (1985). Using interactive computing to expand intelligence testing: A critique and prospectus. Intelligence, 9, 207-236.

Kahneman, D. (1973). Attention and effort. Englewood Cliffs, NJ: Prentice-Hall.

Kogut, J. (1988). Asymptotic distribution of an IRT person fit index. Enschede: University of Twente, Department of Education.

Lehman, E.L. (1959). Testing statistical hypotheses. New York: Wiley.

Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-56.

Levine, M.V., & Drasgow, F. (1988). Optimal appropriateness measurement. Psychometrika, 53, 161-176.

Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics, 4, 269- 290.

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

Matarazzo, J.D. (1983). Computerized psychological testing. Science, 1, 221-323.

McBride, J.R. (1977). Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1, 121-140.

Molenaar, I.W., & Hoijtink, H. (1987). The many null distributions of person fit indices (Heymans Bulletin HB 87-846-EX). Vakgroep S&M FSW, University of Groningen, Netherlands.

Owen, R.J. (1969). A Bayesian approach to tailored testing (Research Report 69-92). Princeton, NJ: Educational Testing Service.

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.

Rudner, L.M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Thissen, D., & Wainer, H. (1983). Confidence envelopes for item response theory (Research Report RR-83-25). Princeton, NJ: Educational Testing Service.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185- 201.

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

Weiss, D.J., & McBride, J.R. (1984). Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 8, 273-285.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.