DOCUMENT RESUME

ED 121 819 TM 005 250

AUTHOR	McBride, James R.; Weiss, David J.
TITLE	Some Properties of a Bayesian Adaptive Ability Testing Strategy.
INSTITUTION	Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO	RR-76-1
PUB DATE	Mar 76
NOTE	44p.
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS	*Ability; Bayesian Statistics; Branching; *Computer Oriented Programs; Correlation; Guessing (Tests); Item Banks; Scores; *Sequential Approach; *Simulation; Test Bias; *Testing
IDENTIFIERS	*Bayesian Adaptive Ability Testing; Tailored Testing

ABSTRACT

Four monte carlo simulation studies of Owen's Bayesian sequential procedure for adaptive mental testing were conducted. Whereas previous simulation studies of this procedure have concentrated on evaluating it in terms of the correlation of its test scores with simulated ability in a normal population, these four studies explored a number of additional properties, both in a normally distributed population and in a distribution-free context. Study 1 replicated previous studies with finite item pools, but examined such properties as the bias of estimate, mean absolute error, and correlation of test length with ability. Studies 2 and 3 examined the same variables in a number of hypothetical infinite item pools, investigating the effects of item discriminating power, guessing, and variable vs. fixed test length. Study 4 investigated some properties of the Bayesian test scores as latent trait estimators, under three different configurations (regressions of item discrimination on item difficulty) of item pools. The properties of interest included the regression of latent trait estimates on actual trait levels, the conditional bias of such estimates, the information curve of the trait estimates, and the relationship of test length to ability level. The results of these studies indicated that the ability estimates derived from the Bayesian test strategy were highly correlated with ability level. However, the ability estimates were also highly correlated with number of items administered, were nonlinearly biased, and provided measurements which were not of equal precision at all levels of ability. (Author)

TM

SOME PROPERTIES OF A BAYESIAN ADAPTIVE ABILITY TESTING STRATEGY

JAMES R. MCBRIDE
AND
DAVID J. WEISS

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

RESEARCH REPORT 76-1

MARCH 1976

Prepared under contract No. N00014-76-C-0243, NR150-382
with the
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 76-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Some Properties of a Bayesian Adaptive Ability Testing Strategy		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) James R. McBride and David J. Weiss		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0243
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, Minnesota 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.:61153N PROJ.:RR042-04 T.A.:RR042-04-01 W.U.:NR150-382
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE March 1976
		13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Portions of this paper were presented at the Spring 1975 meeting of the Psychometric Society, Iowa City, Iowa, April 24, 1975, and the Conference on Computerized Adaptive Testing, Washington, D.C., June 12, 1975.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

testing	sequential testing	programmed testing
ability testing	branched testing	response-contingent testing
computerized testing	individualized testing	automated testing
adaptive testing	tailored testing	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Four monte carlo simulation studies of Owen's Bayesian sequential procedure for adaptive mental testing were conducted. Whereas previous simulation studies of this procedure have concentrated on evaluating it in terms of the correlation of its test scores with simulated ability in a normal population, these four studies explored a number of additional properties, both in a normally distributed population and in a distribution-free context. Study 1 replicated previous studies with finite item pools, but examined such properties as the bias of estimate, mean absolute error, .

and correlation of test length with ability. Studies 2 and 3 examined the same variables in a number of hypothetical infinite item pools, investigating the effects of item discriminating power, guessing, and variable vs. fixed test length. Study 4 investigated some properties of the Bayesian test scores as latent trait estimators, under three different configurations (regressions of item discrimination on item difficulty) of item pools. The properties of interest included the regression of latent trait estimates on actual trait levels, the conditional bias of such estimates, the information curve of the trait estimates, and the relationship of test length to ability level. The results of these studies indicated that the ability estimates derived from the Bayesian test strategy were highly correlated with ability level. However, the ability estimates were also highly correlated with number of items administered, were non-linearly biased and provided measurements which were not of equal precision at all levels of ability.

CONTENTS

SOME PROPERTIES OF A BAYESIAN ADAPTIVE ABILITY TESTING STRATEGY

Adaptive or tailored ability testing subsumes a number of different strategies for adapting the difficulty of test items to the examinee's ability level. All the adaptive testing strategies have as one objective the improvement of the psychometric properties of mental test scores throughout the range of the trait of interest (e.g., ability). This is accomplished by adapting test item difficulty to each examinee's ability, during the test itself. Ideally the adaptive selection and administration of test items would result in each examinee answering only those items which are most informative for his own ability level. Additionally, where items can be answered correctly by random guessing (e.g., multiple-choice items), an optimally efficient adaptive item selection technique would have the effect of equalizing the effect of guessing on test scores throughout the ability range.

The different item selection techniques of the various adaptive testing strategies have been described by Weiss (1974). One of the most elegant of the adaptive strategies is a Bayesian sequential technique proposed by Owen (1969, 1975) and studied empirically by several investigators including Wood (1971), Urry (1971) and Jensema (1972).

Owen's Bayesian Sequential Adaptive Testing Strategy

Owen's technique is a general one for the sequential design and analysis of independent experiments with a dichotomous response. Its application in mental testing is to the problem of estimating ability by means of sequential selection, administration, and scoring of dichotomous test items. The mathematical details of the method arise from latent trait theory, with the item characteristic curves all assumed to take the form of the normal ogive. The properties of the normal ogive item characteristic function and its logistic approximation have been described by Lord & Novick (1968) and Birnbaum (1968), respectively.

Owen's procedure involves the individually tailored sequential design of a test by appropriate choice of available item parameters[1] and estimation of ability (Θ) via a Bayesian-motivated approximation. At each step m in the ability estimation sequence a normal prior distribution on Θ is assumed, with parameters μ_m and σ_m, where m indicates the number of items already administered in the sequence. A test item to be administered at step m+1 is selected so as to minimize a quadratic loss function on Θ. With no guessing (i.e., c_g=0) and the discrimination parameters a_g constant over items, the appropriate item is the available one which minimizes the absolute value of the difference $(b_g-\mu_m)$. With c_g>0 the optimal difference

[1]Each item g can be characterized by three parameters--a_g, b_g, c_g--which are, respectively, the item discriminating power, item difficulty, and item guessing parameter. The guessing parameter, c_g, is simply the probability of answering the item correctly by chance alone.

is somewhat negative; that is, optimal difficulty is somewhat "easier" than examinee's ability.

Following item administration at step $m+1$, the parameters μ_m and σ^2_m of the prior distribution are updated in accord with the examinee's performance on the item. In the case of a correct answer:

$$\mu_{m+1} = E(\theta|1) = \mu_m + (1-c_g)\left(\frac{\sigma^2_m}{\sqrt{\frac{1}{a^2_g} + \sigma^2_m}}\right)\left(\frac{\phi(D)}{c_g + (1-c_g)\ \Phi\ (-D)}\right) \qquad [1]$$

and

$$\sigma^2_{m+1} = \mathrm{var}(\theta|1) = \sigma^2_m \left\{1 - \left(\frac{1-c_g}{1+\frac{1}{a^2_g\sigma^2_m}}\right)\left(\frac{\phi(D)}{A}\right)\left(\frac{(1-c_g)\phi(D)}{A} - D\right)\right\}. \qquad [2]$$

Following a wrong answer:

$$\mu_{m+1} = E(\theta|0) = \mu_m - \left(\frac{\sigma^2_m}{\sqrt{\frac{1}{a^2_g} + \sigma^2_m}}\right)\left(\frac{\phi(D)}{\Phi(D)}\right) \qquad [3]$$

and

$$\sigma^2_{m+1} = \mathrm{var}(\theta|0) = \sigma^2_m \left\{1 - \left(\frac{\phi(D)}{1 + \frac{1}{a^2_g\sigma^2_m}}\right)\left(\frac{\frac{\phi(D)}{\Phi(D)} + D}{\Phi(D)}\right)\right\}. \qquad [4]$$

In Equations 1 through 4 (taken from Owen, 1975)

$\phi(D)$ is the normal probability density function,

$\Phi(D)$ is the cumulative normal distribution function, and

$$D = \frac{b_g - \mu_m}{\sqrt{\frac{1}{a^2_g} + \sigma^2_m}} \qquad [5]$$

$$A = c_g + (1-c_g)\ \Phi\ (-D)\ . \qquad [6]$$

The parameters μ_{m+1} and σ^2_{m+1} of the Bayes posterior distribution on Θ are used as the parameters of the next step's prior. At each step the prior distribution is assumed to be normal. Testing may be terminated when σ^2_m becomes arbitrarily small or when m becomes arbitrarily large, or when some other criterion has been reached. At termination the latest μ_m is the estimator of Θ, and σ^2_m is a measure of the uncertainty of the estimate. Urry (1971) and Jensema (1972, 1974) have interpreted σ^2_m as the squared standard error of estimate (S.E.E.) of Θ_i. Owen (1975) gives a theorem showing that as $m \to \infty$, $\mu_m \to \Theta$; that is, the posterior mean is a consistent estimator of an examinee's ability.

Practically speaking, of course, the number of items administered will never approach infinity; but if the pool of available items is sufficiently large and appropriately constituted, σ^2_m will diminish rapidly, permitting valid estimation of Θ using a small number of items. Urry (1971, 1974) has specified the requirements for a satisfactory item pool for implementing Owen's testing procedure and has shown in computer simulation studies that Owen's sequential test can achieve in 3 to 30 items the validity of a much longer conventional test, with the number of items needed diminishing as item discriminatory power increased.

Urry's (1971, 1974) and Jensema's (1972, 1974) monte carlo simulation studies of Owen's Bayesian testing strategy have evaluated its merit solely in terms of the "fidelity" (or "validity")[2] of the resulting ability estimates and the mean number of items required to achieve any specified value of the fidelity coefficient. Although the fidelity coefficient is of great interest, Lord (1970, p. 152) has pointed out that evaluating an adaptive test by means of a group statistic such as the correlation coefficient presumes some knowledge of the group's distribution on the trait being measured, and ignores information relevant to the accuracy or goodness of the ability estimates at any given level of the trait.

The correlation of test scores with the simulated underlying ability is only one criterion by which to evaluate a proposed adaptive testing strategy. Since the Bayesian sequential test scores are actually estimates of underlying trait level, in the same metric, the accuracy of the estimates is also of interest. "Accuracy" refers to the closeness of the estimates to actual ability; it may vary systematically with ability level. Another interesting property of estimates is bias, or error of central tendency. Two kinds of bias should be of some concern: 1) unconditional bias, or group mean error of estimate; and 2) conditional bias, or mean error of estimate at a given level of the parameter being estimated.

---

[2]By "validity" here is meant the correlation of the ability estimates with actual ability. Green (1975) suggested use of the term "fidelity" in this context to denote validity coefficients obtained from monte carlo simulation studies. Green's convention will be followed here.

Purpose

The purpose of the present paper is to report the results of a series of simulation studies designed to investigate the influence of guessing and item pool characteristics on the bias, accuracy, and other properties of the trait estimates derived from Owen's Bayesian sequential testing strategy.

The studies reported below were motivated by results obtained with live testing of Owen's strategy. Using Owen's testing strategy with 603 college students and a 329-item pool of vocabulary knowledge test items, a correlation of .84 was obtained between estimated ability level and number of test items to termination. Simulation studies then were designed to investigate the influence of item pool characteristics on that unexpectedly large correlation.

The simulation studies reported here were intended to explore both the properties of the Bayesian sequential testing method itself and properties of the resulting ability estimates. The former properties are investigated best by sampling from "populations" of simulated examinees whose distribution on the ability dimension approximates in form and parameters (mean, variance) the population assumed by the testing procedure--here, a normal population with mean 0 and variance 1. The first three studies reported sampled examinees from such a population. These studies were designed to investigate the effects of guessing, of item discriminating power, and of two different test termination criteria on certain group statistics. The independent and dependent variables of interest in each study are described separately below.

The fourth study focused on certain properties of the test scores as estimators of the ability underlying the item responses under varying conditions. This area of inquiry required sampling large numbers of examinees at regular intervals throughout the normal range of the trait. The details of this study are likewise described separately below.

Study 1: An Ideal Item Pool with Variable Test Length

Background and Purpose

Jensema (1972) simulated Bayesian test administration to examinees sampled from a normal [0,1] distribution using two different "ideal" 100-item pools. These pools were "ideal" according to Jensema's prescription that items for use in this testing strategy should have high discriminations and should be rectangularly distributed in their difficulties. The first pool had four items available at each of twenty-five equally spaced difficulty levels in the interval $-2.4 \leq b \leq 2.4$; all items had guessing parameters of $c=.20$ and discriminations of $a=.8$. A second item pool was identical to the first except for the value of the constant discrimination parameter, which was $a=1.60$. The Bayesian test was simulated as proposed by Owen (1969), with the parameters of the initial ability distribution set at [0,1] for each examinee. Testing terminated for each examinee whenever the posterior variance σ^2_m of the ability estimate diminished below a predetermined value or after thirty items, whichever occurred first. Jensema set the critical posterior variance value at .0625, which corresponds to a standard error of estimate of .25, and hence to a fidelity coefficient exceeding .968 (Jensema, 1972, p. 114). Jensema's obtained fidelity coefficients and mean test lengths, obtained from simulations using random samples of 100 examinees, are listed in Table 1.

Table 1
Mean Test Lengths and Obtained Fidelity Coefficients for Two Simulated Bayesian Sequential Tests, Distinguished by their Item Discriminating Power (a)
(from Jensema, 1972)

a	Mean Test Length	Fidelity Coefficient
.80	30*	.93
1.60	17.5	.97

*No tests achieved the posterior variance termination criterion in this condition.

Jensema (1972) did not report, however, some properties of the Bayesian sequential testing procedures which are of practical interest. The purpose of the present study was to replicate Jensema's research with these same two "ideal" item pools, while studying some other properties of the ability estimates in addition to fidelity and mean test length.

Method

Variables. Dependent variables were the individual ability estimates ($\hat{\theta}$) and the number of items (k) required to satisfy the posterior variance termination criterion of $\sigma_m^2 \leq .0625$. Independent variables were the simulated examinees' abilities (θ) and the discriminating power (a=.80 or 1.60) of the items in the simulated item pool.

Examinees' abilities were simulated by computer-generation of 100 random numbers (θ_i) from a normal population with mean 0 and variance 1. The same 100 "examinees" were tested with both item pools.

Item pools. Two 100-item "ideal" item pools were simulated, corresponding to the ones used by Jensema (1972). In each pool there were four items at each of twenty-five difficulty levels (b) equally spaced in the interval $[-2.4<b<+2.4]$. The guessing parameter (c) was constant across items; for both pools, c=.20. The item pool for the first test had a constant discrimination parameter of a=.80 across items; the second pool employed a constant item discrimination parameter equal to a=1.60.

Thus, for each test administration an item pool containing 100 distinct items was simulated; each item g could be characterized by its parameters a_g, b_g, c_g.

Response generation and test administration. Item responses were simulated by calculating, for each item-examinee administration, the probability of a correct response to the item given the simulated ability (θ_i) and the item parameters a_g, b_g, c_g, using equations presented by Betz & Weiss (1974) and Vale & Weiss (1975). This probability $P_g(\theta_i)$ was compared with a random number r_{gi} generated from a uniform distribution in the interval [0,1]. A score of 1 ("correct") for examinee i on item g was assigned if $P_g(\theta_i) \geq r_{gi}$; otherwise a score of 0 was assigned.

Test administration was simulated exactly as proposed by Owen (1969). For each examinee an initial ability $\theta_i=0$ was assumed, and the prior distribution was assumed to be normal [0,1]. The optimal item in the pool was selected based on the item parameters, and its administration to the examinee was simulated. Based on the item score (1 or 0), the parameters (μ_m, σ^2_m) were updated, and another item was selected and administered. This recursive procedure was repeated until 30 items had been taken by the "examinee", or until σ^2_m was smaller than .0625, whichever occurred first. Once any particular item had been taken by the examinee it was not reused At test termination, the examinee's simulated ability (θ_i), the Bayesian estimate ($\hat{\theta}_m$), and the number of items taken (k) were recorded.

Evaluative criteria. For each of the two test administrations, after all 100 examinees' tests were simulated, the following properties of the sequential test were estimated from the data:

a. the bias, or mean algebraic error of the ability estimates;

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i) \quad [7]$$

b. the accuracy, or mean absolute error of the estimates;

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{\theta}_i-\theta_i| \quad [8]$$

c. $r_{\theta k}$, the correlation of test length with ability;
$r_{\hat{\theta} k}$, the correlation of test length with estimated ability;

d. $r_{\theta e}$, the correlation of the algebraic errors of estimate $(\hat{\theta}_i - \theta_i)$ with ability;

$r_{\hat{\theta} e}$, the correlation of $(\hat{\theta}_i - \theta_i)$ with estimated ability;

e. $r_{\hat{\theta} e}$, the fidelity coefficient;

f. the mean, minimum and maximum test length required to achieve the posterior variance termination criterion.

Results

Table 2 contains the results from Study 1. As Table 2 shows, there was positive bias (.06 and .05) in the group scores for both tests, indicating that ability was overestimated, on the average. Mean absolute error was .26 for the a=.80 item pool and .19 for the more discriminating item pool; in these data, then, the more discriminating item pool estimated ability with smaller average error.

Table 2
Properties of the Bayesian Sequential Test for Two Values of Item Discrimination, with Corrected Guessing and Ideal Item Pool

Property	Item Discrimination (a) .80	1.60
Test Length		
Mean	30*	18
Minimum	30	12
Maximum	30	30
Errors of Estimate		
Mean (Bias)	.06	.05
Mean Absolute Error	.26	.19
Correlates		
$r_{\theta e}$	-.35	-.40
$r_{\hat{\theta} e}$	-.07	-.21
$r_{\theta k}$	**	.84
$r_{\hat{\theta} k}$	**	.85
$r_{\theta\hat{\theta}}$	.96	.98

*An arbitrary maximum test length of 30 items was imposed.
**There was no variance on test length in the a_g=.80 test. However θ and $\hat{\theta}$ correlated .81 and .84 with posterior variance.

Mean test length for the a=.80 item pool was 30 items, with no variance, indicating that the posterior variance termination criterion never was reached using this item pool. The higher discriminating pool

(a=1.60) required a mean test length of 18 items, with a range of from 12 to 30. For this item pool test length correlated .84 and .85 with ability and the ability estimator, respectively. This strong positive correlation was essentially the same as was found in the live-testing results. It indicates that despite the "ideal" construction of the item pool, the test required substantially larger numbers of items to achieve the termination criterion as ability increased. (Since there was no variance in test length for the a=.80 item pool, the test length correlations cannot be evaluated under that item pool configuration.)

Errors of estimate ($\hat{\theta}_i - \theta_i$) correlated -.35 and -.40 with ability for the two item pools, which could indicate a tendency to underestimate ability at high levels and to overestimate it at low levels. This, of course, is a phenomenon typical of regression estimates; the Bayesian test scores seem to be acting like regression estimates in this regard. This same tendency was evident to a smaller extent in the correlations between errors and ability estimates ($r_{\hat{\theta}e}$).

The fidelity coefficients ($r_{\theta\hat{\theta}}$) were .96 and .98, respectively, for the a=.80 and a=1.60 item pools. These were slightly higher than those obtained by Jensema (see Table 1). The differences are likely due to random fluctuations resulting from the relatively small sample size of 100 simulated testees (see Betz & Weiss, 1974, pp. 20-21 and 24-25).

Conclusions

The replication of Jensema's study of the Bayesian sequential test using these two item pools corroborated his findings with regard to fidelity and mean test length. The fidelity coefficients obtained in the present study were slightly higher than his, while mean test lengths were almost identical. It seems clear that Owen's adaptive testing procedure has the potential of achieving measurement of high fidelity with relatively short tests. However, the strong correlation between ability and test length suggests a potential problem if the Bayesian test is used in a group of higher ability than is assumed beforehand. Additionally, the overall positive bias of the trait estimates suggests that additional study of the testing procedure is required before its scores are used directly as estimators of ability. However, the generality of the results of Study 1 is limited to "ideal" item pools with rectangular distributions of the difficulty parameters and with the same discrimination and guessing parameters as in the present study.

Study 2: Effects of Guessing and Item Discrimination in a Perfect Item Pool

Background and Purpose

The discovery in Study 1 of positive bias in the Bayesian trait estim^tes, and of a strong positive correlation between ability and test length in the a=1.60 item pool, raises the question of the generalizability of these phenomena. These results might be due to sampling fluctuations, to the specific item parameters employed, to the effects of random guessing, or to characteristics inherent in Owen's sequential testing procedure. Study 2 was designed to test the generality of the results of Study 1.

In Study 2 many sequential tests were simulated by varying the discriminating power of the item pool and the effect of guessing. Further, in order to avoid loss of generality due to a specific range of the distribution of item difficulty values in the item pool, Study 2 simulated a "perfect" item pool--one behaving as though it contained an unlimited number of items at any specifiable difficulty level. The results of Study 2, therefore, should reflect the best attainable results under the Bayesian procedure, given the guessing and discrimination parameters of the items.

To evaluate the effects of guessing on testing strategy characteristics, test administration was simulated under the three different guessing conditions described below--no guessing, uncorrected guessing, and corrected guessing. Under each of these conditions fourteen "perfect" item pools were simulated. These differed from one another only in their item discriminating powers. Thus, fourteen values of a were used; a was constant within any test simulation, but varied across tests. The same properties of the test procedure studied in Study 1 were of interest in Study 2.

Method

Variables. Dependent variables in Study 2 were the same as in Study 1: ability estimates ($\hat{\theta}$) and test length (k). Independent variables were simulated ability (θ), discriminating power of the item pool, and the effect of guessing and of scoring for guessing.

To study the effect of guessing, three different conditions were simulated:

1. No guessing; in the item response model, c was set to 0, and was assumed to be zero in the Bayesian scoring formulae (Equations 1 through 4).

2. Uncorrected guessing; c was set to .20 in the item response model, but was assumed to be zero in the Bayesian scoring formulae.

3. Corrected guessing; c was set to .20 in both the item response model and the Bayesian scoring formulae.

Under each guessing condition, fourteen test administrations were simulated. These differed only in the constant value of the item discriminating powers in the respective item pools. The fourteen values used were a= .5, .6, .7, .8, .9, 1.0, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, and 3.00. For each test administration, the same 100 simulated ability values used in Study 1 constituted the examinee "group".

Item pools. The "perfect" item pools were simulated by calculating, for each examinee after each item response was scored, the optimal difficulty value of the next item, given a_g, c_g and the current ability estimate. This optimal item difficulty was determined using a formula given by Birnbaum (1968, p. 464) for calculating the difficulty level at which maximal item information occurs, given a, c, and assuming that $\hat{\theta}_i = \theta_i$. With a_g constant and when no guessing is assumed ($c_g = 0$ in the scoring formula), the optimal item is one with $b_{m+1} = \hat{\theta}_m$. When guessing is assumed, the optimal difficulty (b_{m+1}) is smaller than $\hat{\theta}_m$, by an amount which is inversely proportional to a_g.

After the "optimal" item difficulty value was calculated, the computer simulation program generated a hypothetical item with that difficulty value, then "administered" it to the examinee. Thus, the hypothetical item pool literally had available an unlimited number of items of any difficulty value specified by the sequential testing procedure.

Response generation and test administration. Item responses were simulated in the same manner described in Study 1. Test administration was identical with Study 1, except for the item difficulty generation procedure. The same posterior variance criterion ($\sigma^2_m \leq .0625$) was used as a test termination rule. Unlike Study 1, test length was free to exceed 30 items; a maximum length of 100 items was imposed. At test termination, ability (θ_i), the ability estimate ($\hat{\theta}_i$), and the number of items administered (k) were recorded for each examinee.

Analysis. A total of 42 test administration conditions were simulated--14 "item pools" under each of the three guessing conditions. For each test administration, the same sequential test properties estimated in Study 1 were estimated: bias, mean absolute error, $r_{\theta k}$, $r_{\hat{\theta} k}$, $r_{\theta e}$, $r_{\hat{\theta} e}$, $r_{\theta\hat{\theta}}$, and the mean and range of test length.

Results

No-guessing condition. As Table 3 shows, test length was constant within item discrimination level under no-guessing, and diminished inversely with level of item discrimination. The posterior variance termination criterion was reached for all examinees using every item pool except the one having a=.50. As a point of comparison with Study 1, test termination was achieved in fewer than 30 items for item pools having $a \geq 1.00$. There was no correlation between test length (k) and θ or $\hat{\theta}$, since there was no variance in test length for any test administration.

The overall bias of estimate under the no-guessing condition was practically zero for all but the highly discriminating item pools (see Table 3 and Figure 1). Mean absolute error was .17 for a=.5 and increased fairly steadily to .22 for the a=3.00 item pool. For the no-guessing condition, then, there is a tendency for the highly discriminating item

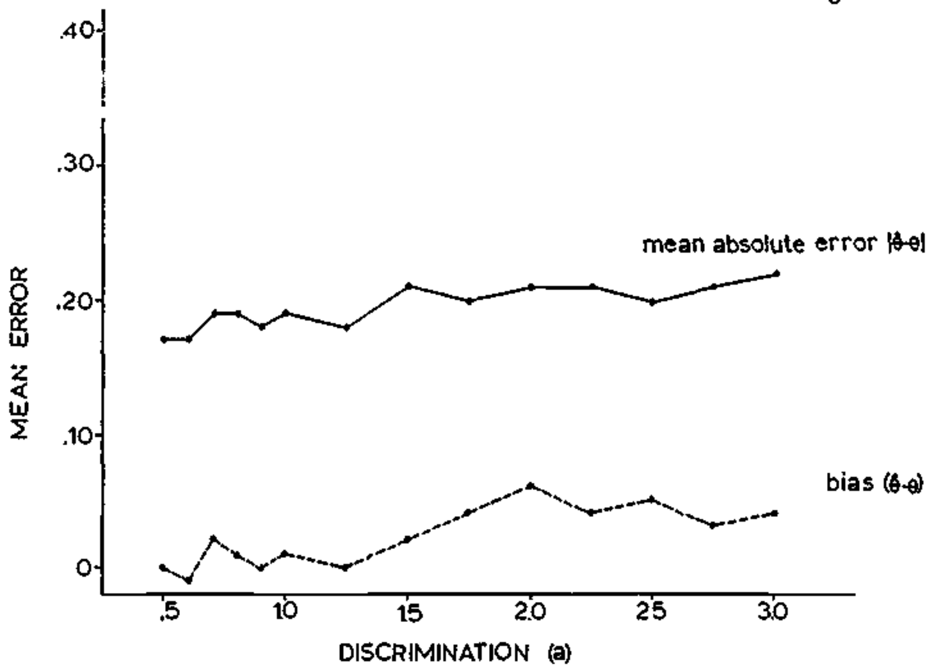
Table 3
Test Length, Mean Errors of Estimate, and Correlates of Ability (θ) and Test Score ($\hat{\theta}$) as a Function of Item Discrimination (a) in the Perfect Item Pool, with No Guessing

Property	Item Discrimination (a)													
	.5	.6	.7	.8	.9	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
Test Length														
Mean	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Minimum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Maximum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Errors of Estimate														
Mean (Bias)	.00	-.01	.02	.01	.00	.01	.00	.02	.04	.06	.04	.05	.03	.04
Mean Absolute Error	.17	.17	.19	.19	.18	.19	.18	.21	.20	.21	.21	.20	.21	.22
Correlates*														
With Error														
$r_{\theta e}$	-.35	-.27	-.31	-.36	-.39	-.35	-.37	-.37	-.30	-.37	-.39	-.36	-.32	-.35
$r_{\hat{\theta} e}$	-.17	-.08	-.10	-.16	-.20	-.15	-.17	-.14	-.07	-.15	-.16	-.14	-.09	-.10
Fidelity (validity)														
$r_{\theta\hat{\theta}}$	.98	.98	.98	.98	.98	.98	.98	.97	.97	.97	.97	.97	.97	.97

*Correlations with test length ($r_{\theta k}$ and $r_{\hat{\theta} k}$) were not computed since test length (k) was constant.

pools to yield *larger* average errors than the moderately discriminating item pools.

Figure 1
Bias and Mean Absolute Error as a Function of Item Discriminations, for the Perfect Item Pool with No Guessing



As in Study 1, errors of estimate ($\hat{\theta}_i - \theta_i$) correlated negatively with θ (-.27 to -.39) and with $\hat{\theta}$ (-.08 to -.20). Again, these correlations suggest a regression effect.

The fidelity coefficients were all .97 or .98, as "predicted" by the posterior variance termination criterion value. Interestingly, the lower fidelity coefficients occurred at the higher item pool discrimination values.

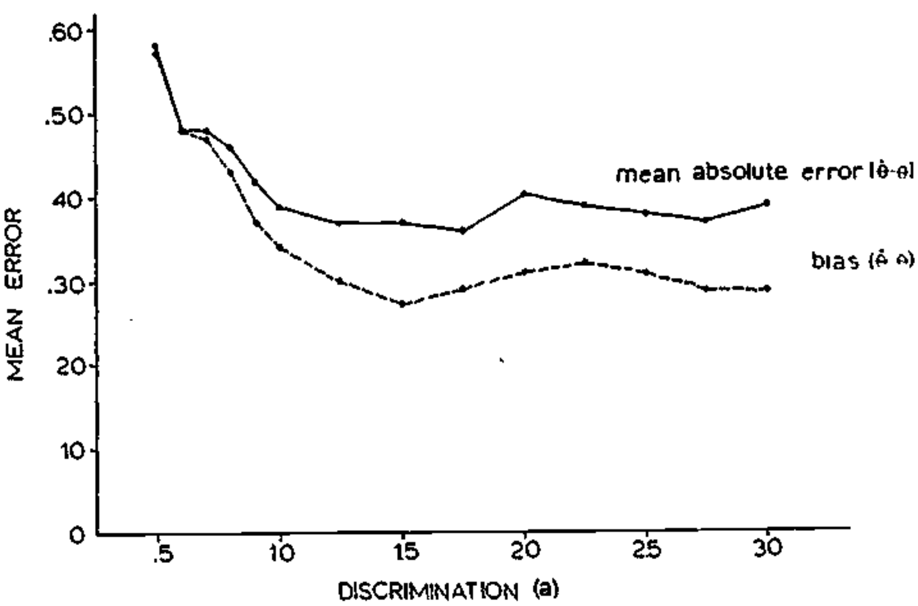
Table 4
Test Length, Mean Errors of Estimate, and Correlates of Ability (θ) and Test Score ($\hat{\theta}$) as a Function of Item Discrimination (a) in the Perfect Item Pool, with Uncorrected Guessing

Property	Item Discrimination (a)													
	.5	.6	.7	.8	.9	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
Test Length														
Mean	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Minimum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Maximum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Errors of Estimate														
Mean (Bias)	.57	.48	.47	.42	.37	.34	.30	.27	.29	.31	.32	.31	.29	.29
Mean Absolute Error	.58	.48	.48	.46	.42	.39	.37	.37	.36	.40	.39	.38	.37	.39
Correlates*														
With Error														
$r_{\theta e}$	-.51	-.46	-.49	-.48	-.48	-.43	-.44	-.36	-.31	-.31	-.32	-.32	-.32	-.32
$r_{\hat{\theta} e}$	-.29	-.23	-.23	-.19	-.20	-.13	-.16	-.04	-.01	[illegible]	.05	.05	.07	.02
Fidelity (validity)														
$r_{\theta\hat{\theta}}$	.97	.97	.96	.95	.95	.95	.96	.94	.95	.93	.93	.93	.92	.91

*Correlations with test length ($r_{\theta k}$ and $r_{\hat{\theta} k}$) were not computed since test length (k) was constant.

<u>Uncorrected-guessing condition</u>. As Table 4 shows, the test length data were identical with those obtained under the no-guessing condition. Table 4 and Figure 2 show that both mean algebraic errors (bias) and absolute errors were quite high (.57, .58) for the a=.50 item pool and decreased as a in-increased, to about a=1.25. For a>1.25 the mean errors seemed to level off, with moderately large values for both bias and absolute error.

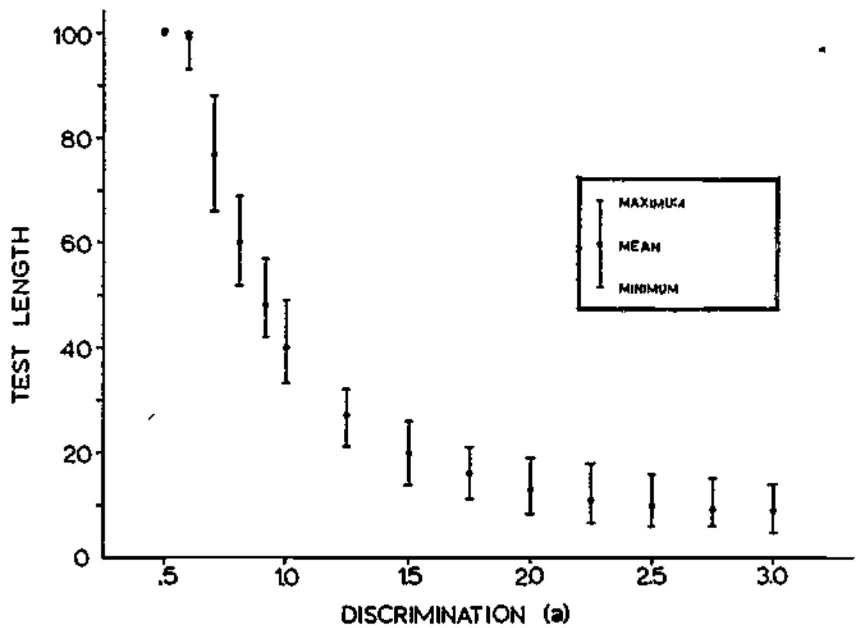
Figure 2
Bias and Mean Absolute Error as a Function of Item Discriminations, for the Perfect Item Pool with Uncorrected Guessing



As before, errors of estimate correlated negatively with ability; the magnitude of the correlations were large for a=.50, then decreased as a increased, until approaching a constant value at a>1.75. Again, these correlations suggest a regression effect. The correlations of errors with ability estimates, $r_{\hat{\theta}e}$, followed a different trend under this condition than was seen previously: $r_{\hat{\theta}e}$ was -.29 for a=.50, then showed a steady algebraic increase with a, to a value of .07 at a=2.75.

Fidelity coefficient values were everywhere lower with uncorrected guessing than with corrected guessing, and *decreased* steadily from .97 to .91 as a increased. As expected, fidelity increased with test length.

Figure 3
Number of Items to Termination, with .20 Guessing



Corrected-guessing condition. Figure 3 graphically depicts test length as a function of item discriminatory power (a). The vertical bars in Figure 3 indicate the range of test length at a given a-level; the dot indicates the mean test length for that level. As Table 5 and Figure 3 show, some variance in test length was present for all a levels except a=.50 (where the termination criterion never was reached). Mean test length to termination varied inversely with item discrimination, as in the other conditions. Even with this perfect item pool, the termination criterion was achieved in fewer than 30 items only for a>1.00.

As Figure 4 shows, the bias of estimate was small but positive under the corrected guessing condition, increasing to meaningful levels only as item pool discrimination exceeded a=2.25. Mean absolute error was almost constant across levels of a.

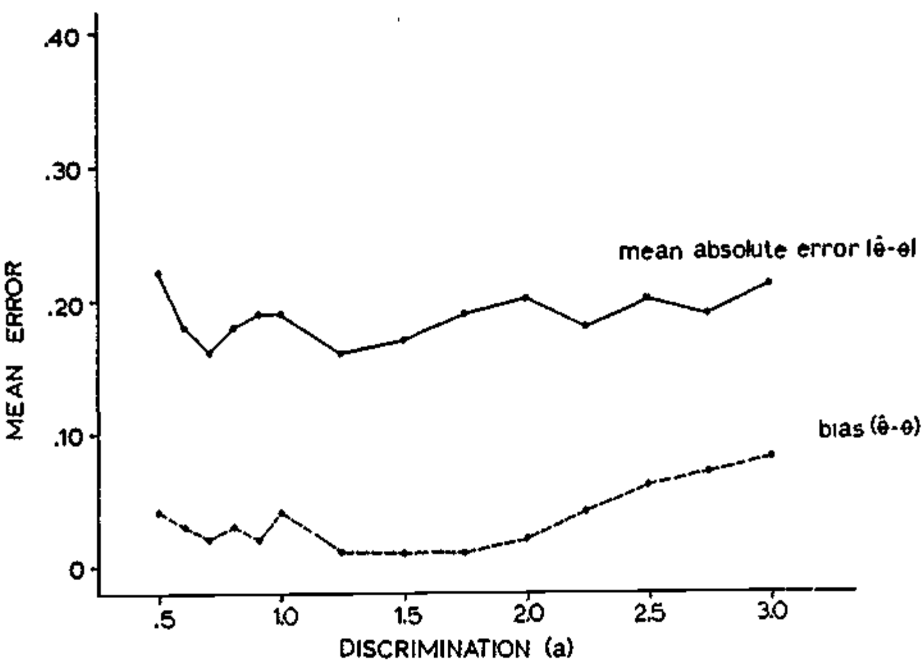
Table 5
Test Length, Mean Errors of Estimate, and Correlates of Ability (Θ) and Test Score ($\hat{\theta}$) as a Function of Item Discrimination (a) in the Perfect Item Pool, with Corrected Guessing

Property	.5	.6	.7	.8	.9	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
	Item Discrimination (a)													
Test Length														
Mean	100	99	77	60	48	40	27	20	16	13	11	10	9	9
Minimum	100	93	66	52	42	33	21	14	11	8	7	6	6	5
Maximum	100	100	88	69	57	49	32	26	21	19	18	16	15	14
Errors of Estimate														
Mean (Bias)	.04	.03	.02	.03	.02	.04	.01	.01	.01	.02	.04	.06	.07	.08
Mean Absolute Error	.22	.18	.16	.18	.19	.19	.16	.17	.19	.20	.18	.20	.19	.21
Correlates														
With Error														
$r_{\theta e}$	-.39	-.36	-.25	-.39	-.42	-.35	-.37	-.37	-.38	-.39	-.25	-.37	-.33	-.33
$r_{\hat{\theta}e}$	-.17	-.18	-.09	-.20	-.23	-.16	-.19	-.18	-.18	-.19	-.14	-.14	-.10	-.08
With Test Length														
$r_{\theta k}$	...*	.54	.80	.78	.78	.81	.81	.82	.85	.88	.85	.88	.90	.88
$r_{\hat{\theta}k}$	...*	.56	.82	.81	.80	.83	.82	.84	.87	.89	.86	.90	.91	.90
Fidelity (validity)														
$r_{\theta\hat{\theta}}$	.97	.98	.99	.98	.98	.98	.98	.98	.98	.98	.98	.97	.97	.97

*Correlations not computed since test length (k) was constant.

As was seen in Study 1, test length correlated strongly with ability (and ability estimates) where it was free to vary (Table 5). Since test termination takes place only after a specified reduction of the posterior variance has occurred, the large positive $r_{\theta k}$ correlations indicate that the rate of posterior variance reduction is a function of ability level, with more rapid reduction taking place as ability (Θ) decreases.

Figure 4
Bias and Mean Absolute Errot as a Function of Item Discriminations, for the Perfect Item Pool with Corrected Guessing



As seen under the other conditions, Table 5 shows that errors of estimate correlated negatively (-.25 to -.42) with ability and with ability estimates (-.09 to -.23). As in the no-guessing condition, all fidelity coefficients were .97 or .98, with the lower value occurring at the higher item discrimination levels.

Conclusions

Study 2 supports the findings of Study 1 and extends them somewhat. As in Study 1, the Bayesian testing strategy resulted in very high fidelity coefficients with relatively short tests, provided the item discriminating powers were 1.0 or greater. The Study 1 finding of positive overall bias of estimate was corroborated here: Only one of the forty-two bias estimates was negative. Especially noteworthy was the effect of uncorrected guessing on both the ability estimates and the fidelity coefficients: Bias was severe, and fidelity actually decreased as discriminating power increased.

Under the corrected-guessing condition, the finding of a strong positive correlation between test length and Θ or $\hat{\Theta}$ was replicated consistently. It is important to note that this condition was obtained under conditions of a "perfect" item pool; this implies that the high correlation does not result from inadequacies of the item pool. Since there was no variance in test length when no guessing was assumed (i.e., for the no-guessing and uncorrected-guessing conditions), the phenomenon would seem to be due to the scoring formulae in some way. The phenomenon by itself is of little concern unless it results in different measurement properties at different levels of ability. This may be the case; some of the properties of the sequential test seem to improve with test length. If test length is consistently greater as ability increases, then the test may be measuring less well as ability decreases, due simply to the effects of test length.

Study 3: Effects of Fixed Test Length

Background and Purpose

The results of Study 2 make it obvious that with guessing a factor, test length increases with ability level when the posterior variance criterion is used to terminate testing. It was suggested that some measurement properties of the test may suffer as a consequence. Two properties which seem to be affected adversely by short test length are bias and mean absolute error, both of which increased as item discrimination became very high (and test length very short) in the no-guessing and corrected-guessing conditions (see Tables 3 and 5). Another property which should be adversely affected by very short test lengths is fidelity. Study 2 noted a small but consistent decline in fidelity at the very high discrimination levels (see Tables 3, 4 and 5). Additionally, Jensema (1972) noted a similar phenomenon, which he termed "correlation drop-off".

This study explored the effect of administering the same number of items to all examinees, on the same properties which were of interest in Studies 1 and 2. This was done by means of simulating fixed-length Bayesian tests for the corrected-guessing condition, under various item discrimination levels. To avoid loss of generality, the "perfect" item pool was again employed.

Method

Variables. Dependent variables were the ability estimates ($\hat{\theta}$) and the posterior variance (σ_k^2) after a fixed number (k) of items had been administered. Independent variables were simulated ability (θ) and item discriminating power. Nine levels of discriminating power were studied: a_g=.6, .8, 1.0, 1.25, 1.50, 1.75, 2.0, 2.5, 3.0. Examinees were the same 100 simulated ability values (θ_i, i=1, 2, ... 100) used in Studies 1 and 2.

Item pools. "Perfect" item pools were simulated, as described in Study 2; i.e., the locally optimum item difficulty was calculated after each item response, and an item having that difficulty level was artificially generated and administered.

Response generation and test administration. Item responses were simulated in the same manner as in Studies 1 and 2. Test administration was identical with Study 2, except that all "examinees" were administered 30 items. After 30 items, the individual ability (θ_i), the estimate ($\hat{\theta}_i$), and the posterior variance (σ_{30}^2) were recorded for each examinee.

Analysis. A total of nine test administrations were simulated (one at each item discrimination level). For each administration these sequential test properties were estimated as described in Study 1: bias, mean absolute error, $r_{\theta e}$, $r_{\hat{\theta} e}$, and $r_{\theta\hat{\theta}}$. Additionally, for each administration, the correlations of the posterior variance with θ and $\hat{\theta}$ were calculated.

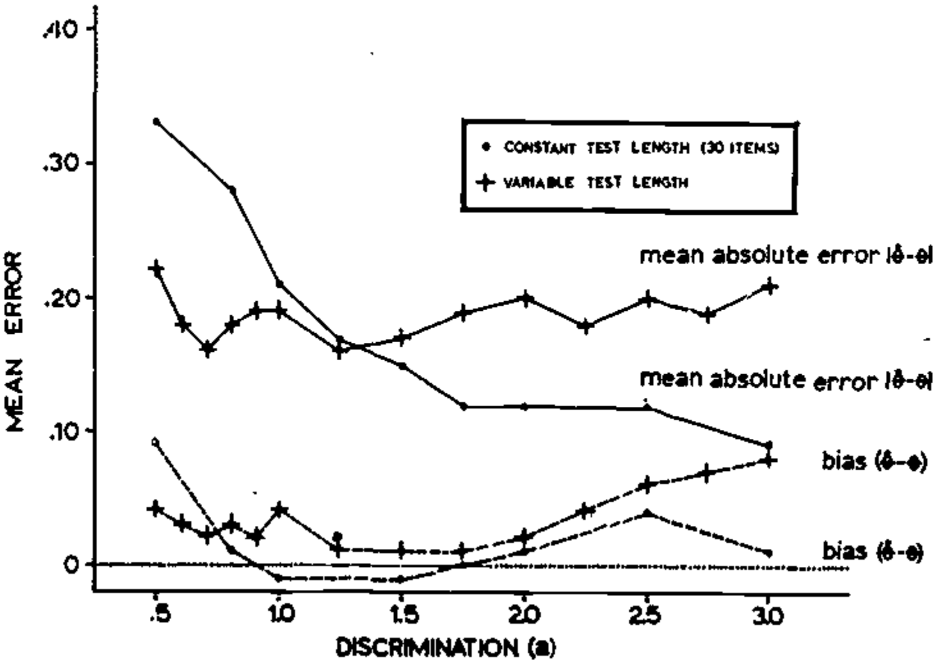
Results

Table 6 and Figure 5 contain the results of Study 3. To facilitate comparing the 30-item test length with the posterior variance termination criterion, comparable data from Study 2 are included in Figure 5.

As Figure 5 shows, the overall bias of estimate was virtually zero in all item pools, except for the a=.60 and a=2.5 item pools. Mean absolute error decreased steadily as a function of a, and was lower for fixed test length than for the variable test length conditions for all discriminations larger than a=1.50. As in Studies 1 and 2, error ($\hat{\theta}_i-\theta_i$) correlated negatively with θ and $\hat{\theta}$, suggesting a regression effect.

As Table 6 shows, the posterior variance correlated positively with θ and $\hat{\theta}$, with the magnitude of the correlation generally diminishing as a increased (e.g., $r_{\theta\sigma_{30}^2}$ was .86 for a=.6, and .74 for a=3.0). This trend corresponds to the one seen in Studies 1 and 2--test length correlates strongly with ability when posterior variance is held constant.

Figure 5
Mean Absolute Error and Bias for Two Different
Test Termination Criteria



The fidelity coefficients increased with the item discriminating power, from .93 at a=.60 to .99 at a=1.5 and higher.

Table 6
Errors of Estimate and Correlates of the Bayesian Sequential Test Ability Estimates as a Function of Item Discrimination, for 30-Item Test Length and Corrected Guessing, with Perfect Item Pool

Property	Item Discrimination (a) .6	.8	1.0	1.25	1.5	1.75	2.0	2.5	2.75
Errors of Estimate									
Mean (Bias)	.09	.01	-.01	.02	-.01	.00	.01	.04	.01
Mean Absolute Error	.33	.28	.21	.17	.15	.12	.12	.12	.09
Correlates									
With Error									
$r_{\theta e}$	-.41	-.30	-.36	-.34	-.40	-.32	.32	-.51	-.36
$r_{\hat{\theta} e}$	-.04	.01	-.13	-.15	-.24	-.19	-.18	-.36	-.23
With Posterior Variance									
$r_{\theta\sigma^2_m}$	.86	.85	.89	.81	.82	.77	.69	.76	.74
$r_{\hat{\theta}\sigma^2_m}$	.93	.90	.90	.84	.82	.79	.69	.72	.73
Fidelity									
$r_{\hat{\theta}\theta}$	.93	.95	.97	.98	.99	.99	.99	.99	.99

Conclusions

It is apparent that some improvement in the properties of the Bayesian testing procedure can be realized by setting test length constant, provided that item discriminatory power is sufficiently high (e.g., greater than $a=1.5$). Bias seems to be diminished, and absolute error decreases as discrimination increases.

Study 4: Effects of Ability Level and Item Pool Configuration

Background and Purpose

Simulation studies of Owen's Bayesian sequential test procedure typically have concentrated their attention on group statistics. For example, Urry (1971, 1974) and Jensema (1972, 1974) evaluate their results in terms of fidelity coefficients and mean test length (using a posterior variance termination criterion). Studies 1, 2, and 3 above have extended Urry's and Jensema's work by examining additional properties of the sequential testing procedure, but they also concentrate on group statistics. With any group statistic, such as a fidelity coefficient, a bias estimate, or a mean test length, there is a lack of invariance across groups. A change in the shape of the distribution, or the central tendency and variability, may alter the magnitude of the group statistic markedly. Therefore, some distribution-free methods for evaluating the Bayesian sequential adaptive test are needed. One general method for this is to examine characteristics of the test as a function of ability level.

Given that some properties are to be evaluated as a function of ability level, it is necessary to select the properties of interest. The results of Studies 1, 2, and 3 suggest some characteristics of Owen's procedure which bear further investigation. For instance, there was a tendency in the preceding studies for positive bias to occur, i.e., for the group average ability estimates to be larger than the average ability. Additionally, there was consistently a moderate negative correlation between ability and the errors of estimate, indicating a regression effect. The negative correlation between the estimates themselves and their error further suggests that the regression may be non-linear. The strong positive correlation between test length and ability indicates that the posterior variance estimate is being reduced more rapidly at low ability levels than at high ones, despite the use of the "perfect" item pools and the presence of constant item discrimination across all difficulty levels.

Based on the findings of Studies 1, 2, and 3, the present study examined appropriate properties of the Bayesian sequential testing strategy as a function of ability level. These properties include the form of the regression of ability estimates on Θ, the conditional bias of the ability estimates, and mean test length. In addition, this study included estimation of the "information" (Birnbaum, 1968) in the Bayesian test ability estimates at various levels of ability.

In addition to estimating the regression, bias and information in the Bayesian test scores as a function of ability, this study examined the effect which different item pool "configurations" might have on these properties. Item pool configuration here refers to the regression of item discrimination (a) values on the item difficulty (b) values in the item pool. Studies 1, 2, and 3 above, and all previous research using "ideal" item pools, have simulated item pools in which a was constant across items or in which a was statistically independent of b. The presence of no statistical association between a and b implies that the same *item* information (Birnbaum, 1968, p. 449) is available at all levels of item difficulty. On the other hand, if there is a statistical relationship between the discrimination and difficulty values of the items in a given item pool, there will be more information available in some ranges of the ability continuum than there is in others.

Although in theory it is desirable for adaptive testing to assemble an item pool having equally discriminating items at all the difficulty levels represented, in practice this has not always been achieved. For instance, the 58-item pool used by Jensema (1972) to simulate adaptive testing based on some items from the Washington Pre-College examinations had very highly discriminating items in its upper difficulty ranges and low-to-moderately discriminating items in the easy range of difficulty. Similarly, Lord (1974) reported that the discrimination parameters of his item pool correlated positively with the difficulty parameters. Practical implementations of adaptive testing are likely to use item pools in which the configuration of the item parameters is less than ideal. Therefore, the effects of different item pool configurations on the psychometric characteristics of the test scores (or trait estimates) need to be investigated.

This study investigated three different configurations of the item pools. Each configuration was characterized by a different slope of the regression of item discrimination parameters on item difficulty, which in turn can be characterized approximately in terms of the correlation, r_{ab}, between item discriminating power and difficulty. Identical test simulation studies were conducted under all three configurations in order to evaluate any differential effects.

Method

Variables. Dependent variables were the ability estimates ($\hat{\theta}$) and the number of items (k) required to satisfy the test termination criterion. Independent variables were the simulated examinees' abilities (θ_i) and the configuration of the simulated item pool. Examinees' abilities for each test administration were simulated by 3100 values of θ_i, 100 at each of 31 equally spaced levels in the interval $[-3.0 \leq \theta \leq +3.0]$. This examinee distribution was used because of the need for relatively large numbers of observations at each level of θ in order to estimate accurately the regression of ability estimates on ability, the conditional bias, and the information curves.

Item pools. Three "perfect" item pools were simulated--one for each configuration. The three configurations studied included one with a moderate positive correlation of a with b (referred to hereinafter as $r_{ab}+$), one with a moderate negative correlation ($r_{ab}-$), and one with no correlation ($r_{ab}0$). The $r_{ab}+$ configuration favored the more difficult items with higher discriminating powers, the $r_{ab}-$ configuration favored the easier items, and the $r_{ab}0$ configuration favored no difficulty levels.

As in Studies 2 and 3, after each item response the optimal difficulty of the next item to administer was calculated, and an item having that difficulty value was artificially generated and administered. In the previous studies, the optimal difficulty calculation was based on the guessing parameter (c) and on the constant discrimination parameter (a) of the items in the pool. In this study, the same calculation was based on the *mean* item discrimination parameter ($\overline{a}$), which was 1.25 for all configurations. In all cases, c was .20.

The item pool configuration was simulated by:

1. Selecting the appropriate b_g for the next item from the perfect item pool as though all a_g were equal to $\overline{a}_g$; call this $b^*_g=(b_g|\hat{\theta}_m, \overline{a}_g)$;

2. Calculating a conditional a_g value from a linear transform of b^*_g

$$a_g|b^*_g=r_{ab}\left(\frac{S.D._A}{S.D._B}\right)b^*+\overline{a}_g \qquad [9]$$

where $S.D._A$ is the standard deviation of the a_g parameters in the simulated pool;

$S.D._B$ is the standard deviation of the b_g parameters in the simulated pool;

a_g, b^*_g, r_{ab}, $\overline{a}_g$ are as previously defined;

3. Adding an error component, e_g, to the approximate a_g, so that for each item administered $a^*_g=a_g|b^*_g+e_g$ where a^*_g is the simulated discriminating power of the item;

$a_g|b^*_g$ is the approximate discrimination defined above;

e_g is a random number from a normal $[0, \sigma^2_e]$ population, such that

$$\sigma_e = \sqrt{\sigma^2_e} = S.D._A\ (1-r^2_{ab})^{1/2}. \qquad [10]$$

4. Setting a^*_g equal to .80 whenever it would otherwise have a lower value.

Response generation and test administration. Item responses were simulated in the same manner described in Study 1. Test administration was identical with Study 1. A posterior variance termination criterion of $\sigma^2_m \leq .0625$ was used, with an arbitrary maximum test length of 30 items. The corrected-guessing condition was used. At termination, the ability (θ_i), its estimate $(\hat{\theta}_i)$, and the number of items administered (k) were recorded for each examinee.

Analysis. For each of the three simulated test administrations, the following properties of the sequential test were estimated from the 100 observations at each separate ability level (θ_i):

a. the conditional mean, $\bar{\hat{\theta}}_i | \theta_i = \frac{1}{100}\Sigma\hat{\theta}_i$ [11]

b. the conditional variance, $\sigma^2_{\hat{\theta}_i|\theta_i} = \frac{1}{100}\Sigma(\hat{\theta}_i - \bar{\hat{\theta}}_i)$ [12]

c. the conditional bias, $b_i|\theta_i = \bar{\hat{\theta}}_i - \theta_i$ [13]

d. the conditional mean test length, $\bar{k}|\theta_i$.

The regression of the trait estimates $(\hat{\theta})$ on ability (θ) was estimated by fitting a third degree polynomial to the 31 conditional means, using a least squares method. The regressions of bias and test length on θ were estimated graphically.

The information in a set of test scores (x) can be defined as

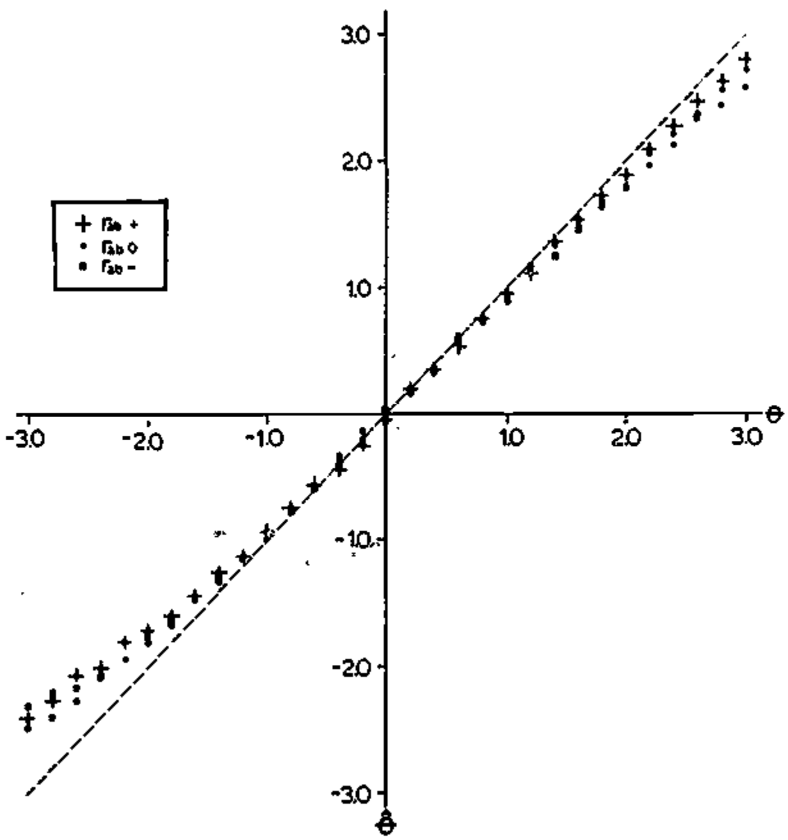
$$I_x(\theta) = \left[\frac{\frac{\partial}{\partial\theta}(E(x|\theta)}{\sigma_{x|\theta}}\right]^2 . \qquad [14]$$

The "information" value of test scores at any level of ability is an index of the usefulness of those scores for discriminating among examinees in the vicinity of that level. A zero information value indicates that the test scores are useless for making discriminations about a given point; an infinite information value indicates that error-free discriminations can be made about that point on the basis of the test scores. Any value between the two extremes has implications for the probability of making Type I and Type II errors in classifying persons above or below the point in question.

The numerator in Equation 14 is the first partial derivative of the function describing the regression of test scores (x) on the trait (Θ). The denominator in Equation 14 is the conditional standard deviation of the scores. The regression of test scores on Θ can be approximated from empirical data, if the scores (x) and the latent trait values (0) are known.

Since the Bayesian trait estimates (Θ) can be treated as test scores, the numerator of the information function can be evaluated at any point (Θ') from the slope of the equation for the regression of $\hat{\Theta}$ on Θ. That equation was calculated from the simulation data as described above. In estimating the information curves, the first partial derivative (i.e., the slope) of that polynomial equation was evaluated at each of the 31 Θ points used in the study. The denominator of the information function at each of the same 31 points was estimated by the square root of the conditional variance of the trait estimates at that point.

Figure 6
Mean Estimated Ability ($\hat{\theta}$) at 31 Ability Points ($\hat{\theta}$)
for the Simulated Bayesian Sequential Test under
Three Item Pool Configurations



Thus for each of 31 points Θ', the information at that point, $I_{\hat{\Theta}}(\Theta')$ was estimated from the test simulation data, as

$$I_{\hat{\theta}}(\theta') = \left[\frac{\frac{\partial}{\partial\theta'}E(\hat{\theta}|\theta')}{\hat{\sigma}_{\hat{\theta}|\theta'}}\right]^2 \qquad [15]$$

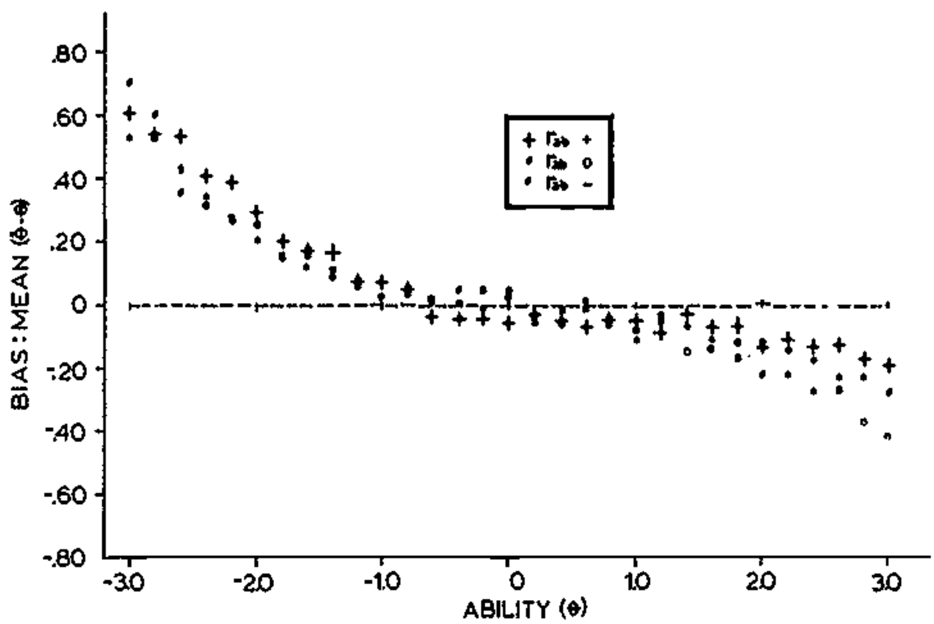
where $E(\hat{\theta}|\theta')$ is the third degree polynomial regression fitted to the 31 test score means

$\sigma(\hat{\theta}|\theta')$ is the square root of the observed variance of the 100 test scores at θ'.

Results

Regression of $\hat{\theta}$ on θ. Figure 6 is a plot of the observed mean ability estimates ($\hat{\theta}$) as a function of actual trait level (θ) differentiated by item pool configuration; Appendix Table A-1 shows the numerical values of these means. For each configuration, then, Figure 6 contains the graphic empirical approximation of the regression of $\hat{\theta}$ on θ. The values for each item pool configuration form an essentially linear plot for levels of θ between +1 and -1, with a tendency toward departure from linearity for values of θ larger than +1 and smaller than -1. High abilities are underestimated; low abilities are overestimated. The exaggeration of this effect seems strongest for the r_{ab}- configuration, in which the average item discrimination increased as the ability estimates decreased.

Figure 7
Mean Error of Estimate ($\hat{\theta}$-θ) at 31 Ability Points (θ)
for the Simulated Bayesian Sequential Test under
Three Item Pool Configurations



Bias. Figure 7 contains the plot of conditional bias (mean ($\hat{\theta}$-θ)) on ability (numerical values are in Appendix Table A-1 as e). For each configuration, the curve described by these data is non-linear. As Figure 6 showed indirectly, the conditional bias for all three configurations was close to zero for $-1 \leq \Theta \leq 1$, but it increased with increases in absolute values of Θ elsewhere. A strong tendency to underestimate high Θ was present in all three configurations, and was severe for $r_{ab}-$, for which the bias was $-.43$ at $\Theta=3.0$. The tendency to overestimate low Θ was even more pronounced, and was severe for all three item pool configurations. For the $r_{ab}0$ configuration the conditional bias at $\Theta=-3$ was .53; for $r_{ab}-$ the bias at the same point was .61. If the Θ metric is expressed in population standard deviation units, then, the Bayesian sequential test estimates may typically err by one-half standard deviation unit at low extremes of the ability range and by a lesser but still significant amount at the high extremes. Furthermore, this tendency is systematically affected by the configuration of the item pool.

Figure 8
Mean Number of Items to Termination (Test Length) at 31
Ability Points (Θ) for the Simulated Bayesian Sequential
Test under Three Item Pool Configurations

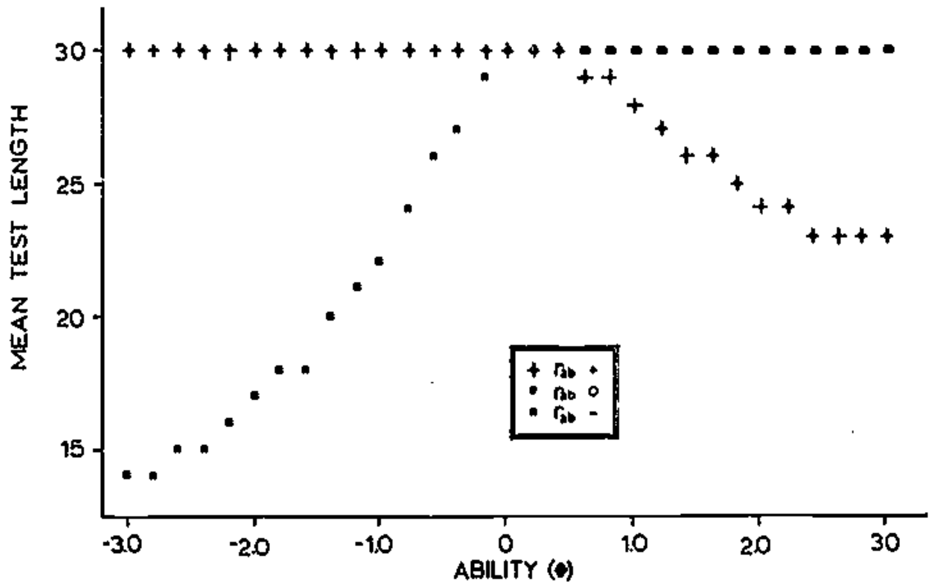
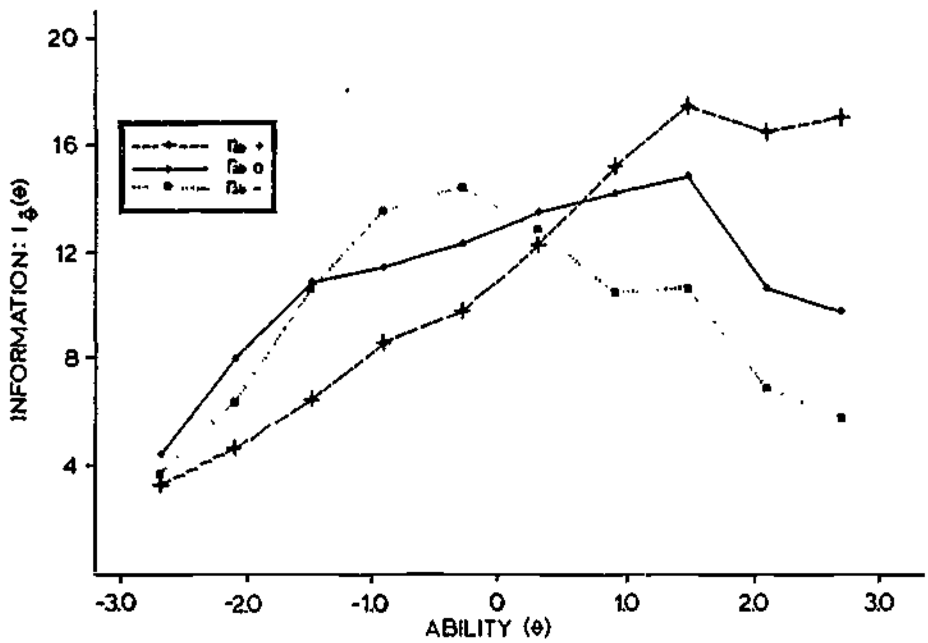


Figure 8 contains plots of mean test length as a function of ability level for each item pool configuration (numerical values are in Appendix Table A-1). For the $r_{ab}0$ configuration, test length was constant at 30 items, the arbitrary maximum. For $r_{sb}+$, where the most discriminating items were available at the higher difficulty levels, test length was constant at 30 items for Θ levels less than .6, then declined gradually to a mean of 23 items at $\Theta=3$. The $r_{ab}-$ configuration, which had higher item discrimination at the lower difficulty levels, showed a trend opposite that for $r_{ab}+$. For

r_{ab}-, test length increased rapidly with Θ from a mean of 14 items at Θ=-3, to 30 items at Θ=0; for all Θ greater than zero, the test length was 30 items, the arbitrary maximum.

Figure 8 illustrates two interesting trends. First, not only did the r_{ab}- configuration use fewer items than the others, but the *rate* of increase as Θ increased is noticeably steeper than the rate of decline in test length for r_{ab}+. Second, for r_{ab}+, which required the fewest items at high Θ levels, bias (see Figure 7) was least pronounced at high Θ levels; yet for r_{ab}-, which required fewest items at low Θ levels, there is no apparent advantage at those levels in terms of bias.

Figure 9
Smoothed Information Curves for the Bayesian Sequential Test under Three Different Item Pool Configurations



Information. Figure 9 contains smoothed information curves for the three item pool configurations. (Numerical values of the estimated slopes, conditional standard deviations, and information values at each of the 31 Θ levels are shown in Appendix Table A-2.) For the r_{ab}0 configuration the information curve shown in Figure 9 is convex, reaching its maximum height very near Θ=0; the curve slopes gradually downward as Θ increases above 0, and more rapidly downward as Θ decreases from 0. At Θ=-3 the information curve is quite low, indicating that despite the availability of test items at all difficulty levels, the test scores will discriminate very poorly in the low ability ranges.

For the $r_{ab}+$ configuration the information value at $\theta=-3$ is even lower, but it increases steadily--almost linearly--with θ. The $r_{ab}+$ information curve surpasses that of $r_{ab}0$ at $\theta\geq+1$, as expected from the availability of more discriminating items in the higher difficulty ranges. For the $r_{ab}-$ configuration, which had its lowest item discriminations in the higher difficulty ranges, the information curve is quite low at high ability levels, and it increases steadily as θ decreases, to about $\theta=0$. Surprisingly, the information curve thereafter decreases with θ, reaching its lowest point at $\theta=-3$. This is a striking result in view of the availability of more discriminating items at low θ levels for the $r_{ab}-$ item pool. It can be partly, but not entirely, accounted for by the shorter test lengths seen for the $r_{ab}-$ configuration at the low ability levels.

General Summary and Conclusions

Previous research (e.g., Urry, 1971, 1974; Jensema, 1972) has shown that Owen's Bayesian sequential approach to adaptive testing has the potential of achieving very high correlations between ability level and ability estimate concomitant with a significant savings in test length, compared to conventional testing procedures. In order for this potential to be realized, a relatively large item pool was required, with highly discriminating items ($a>.80$) rectangularly distributed on the difficulty continuum (Urry, 1974). Study 1 corroborated the findings of Urry and Jensema in terms of test length and values of the fidelity coefficients. At the same time Study 1 revealed an overall tendency for the Bayesian trait estimators to overestimate group mean ability level. Also, the results of Study 1 corroborated the finding in live-testing that with Owen's strategy test length covaries positively with ability level.

The results of Study 1 were not definitive, partly because finite item pools were employed. Study 2 overcame the specificity of Study 1 by introducing the use of a "perfect" (or infinite) item pool, having unlimited numbers of independent items at any difficulty level. At the same time, Study 2 varied the values of the guessing parameter.

The results of Study 2 suggest that the bias problem seen in Study 1 may be largely a result of guessing; under the no-guessing condition bias was virtually zero, except for the very highly discriminating item pools. This relationship was confounded with test length, however, since the highly discriminating item pools reached the test termination criterion in a very small number of items (e.g., 5 items at $a=3.00$). Under the corrected-guessing condition, bias was consistently positive, and increased as item discriminations increased and mean test length became very short. Under the uncorrected-guessing condition, both bias and mean absolute error were pronounced.

The high correlation between test length and ability level was consistently present in Study 2 under the corrected-guessing condition. Under no-guessing and uncorrected-guessing, however, there was no such correlation because there was no variance in test length within a test. Under the latter conditions, test length varied only across tests--i.e., as a function of item discriminating power.

In terms of fidelity coefficients, there was no appreciable difference between those obtained under no-guessing and under corrected-guessing, given the common termination criterion. Under uncorrected-guessing, however, there was some loss of fidelity as test length decreased. It should be noted that the uncorrected-guessing condition was tantamount to assuming an inappropriate item response model. The result of using the inappropriate model to estimate ability and to select items sequentially was to introduce large errors of estimate and some loss of fidelity.

The observation that bias, absolute error, and fidelity seemed to be adversely affected by the short test lengths typical of highly discriminating item pools led to using a fixed 30-item test length in Study 3. The results confirmed the hypothesis that some undesirable psychometric properties may accompany the use of very highly discriminating item pools if the posterior variance criterion is used to terminate testing. When test length remained constant, bias was virtually zero and absolute error diminished steadily as item discrimination increased.

The interrelationships of test length, item discrimination, bias, and absolute error would be a fruitful avenue for further research. If the interdependencies were understood it would be possible for a test user to control error magnitudes by appropriate choice of test length, given knowledge of the parameters of the items in the item pool.

Study 4 investigated some of the characteristics studied earlier but as a function of trait level. The curvilinear regression of the latent trait estimators on trait level illustrates the conservative nature of Bayes estimators. Fairly accurate estimation is achieved in the vicinity of the assumed prior mean, at the expense of accuracy in the extremes. In a sense, the Bayesian procedure gives little "credence" to extreme trait values; this conservatism results in a consistent tendency to underestimate high trait level values and to overestimate low ones. With guessing present the overestimation problem becomes accentuated. This alone may be sufficient to explain the positive bias seen in Studies 1 and 2: The overestimates tend to be of larger magnitude than the underestimates, resulting in an overall tendency towards overestimation.

More significant than the direction of the conditional bias is its form. Under all three item pool configurations in Study 4, the bias curves were non-linear. In ability testing, bias is not usually of concern as long as it is constant or linear in the parameter being estimated (Lord, 1970, p. 153), since these two cases imply a linear relationship between test scores and trait level parameters. Non-linear bias, on the other hand, implies a non-linear relationship, which in turn adversely affects the utility of the test scores. Other things being equal (e.g., the conditional variances of the test scores), if the regression of test scores on trait level is non-linear, the scores will make better discriminations at some trait levels than at others.

That this is the case with the scores resulting from Bayesian test administration is evident in the information curves estimated from the data. Although adaptive testing has the potential to result in equi-discriminating ability estimates, the Bayesian sequential adaptive test has failed to achieve this goal under the conditions simulated in Study 4. Under each item pool configuration, some region of the ability continuum had considerably higher levels of information under any configuration. Even under the r_{ab}- configuration, where the best discriminating items were available in the lowest difficulty regions, the information curve was very low in the low ability region.

Lord (1970, p. 152) indicated that evaluating an adaptive test by means of a group statistic (such as the fidelity coefficient, $r_{\hat{\theta}\theta}$) presumes some knowledge of the group's distribution on the trait being measured, and ignores information relevant to the accuracy of trait estimates at any one level of the trait. The validity of the Bayesian sequential test trait estimates, as the results show, was quite high under the conditions used in these simulation studies. The accuracy of the estimates was also favorable in what corresponds to the middle ranges of a normal distribution on θ, but was found to be less favorable in the extremes, especially the lower extreme. Similarly, the information curves of the trait estimates showed that the effectiveness of measurement under the Bayesian testing procedure varied systematically as a function of the configuration of the item parameters constituting the item pool, but in all three configurations measurement effectiveness was very low in the low ranges of the trait.

The observed loss of accuracy and information in the extremes of the "typical" range of θ are disturbing, since a major advantage of adaptive testing over conventional testing is the former's supposed potential for superior measurement accuracy and effectiveness in those extremes. The data of this series of studies show that with the exception of the r_{ab}+ configuration, the adaptive test scores behave much like conventional test scores, at least in terms of the shapes of their information curves. The utility of the Bayesian adaptive testing strategy may be diminished by results like those reported for Study 4, if they prove to be general.

The problems of bias which is non-linear in θ, and of convex information curves as observed in Study 4, have causes which may be amenable to improvement. Central to both problems is the effect of guessing, which generally operates to reduce measurement efficiency at all trait levels, and especially at low trait levels. Also at the core of the problems is the Bayesian procedure itself. As was pointed out earlier, the Bayesian trait estimates behave like regression estimates. Extreme values of θ are systematically regressed toward the initial prior estimate; the assumption of a normal prior distribution of θ ensures this tendency. On the average, the more extreme θ is for any individual, the larger will be the regression effect. Recall that the item selection procedure selects an item with difficulty somewhat easier than the current θ estimate. But for high θ the current estimate is almost always too low.

Hence the difficulty of the selected item will almost always be too easy for extremely able examinees. Cumulated over 30 items, for example, there will be several effects of this inappropriate item selection:

1. Mean proportion correct will tend to increase as a function of θ, despite the implicit attempt of the tailoring procedure to make it constant at all levels of θ;

2. θ will tend to be underestimated for high θ due to the inappropriate difficulty of the test items administered;

3. Information loss will occur at high θ due to the shallowing slope of the regression of $\hat{\theta}$ on θ.

For low θ the initial prior is an overestimate. Hence the first item selected will generally be too difficult, yet the examinee has a chance of answering it correctly by guessing. A correct answer, of course, will cause an increase in $\hat{\theta}$ and thus result in another inappropriate choice of item difficulty. Furthermore, as Samejima (1973) has shown, when guessing is a factor there may actually be negative information in a correct response to an item whose difficulty exceeds an examinee's actual trait level by a fairly small increment. Thus it appears that in Owen's Bayesian strategy, testees in the low extremes of θ are rather consistently being administered overly difficult items with several systematic results:

1. Mean proportion correct tends to decrease with θ despite the tailoring process;

2. Posterior variance reduction tends to be more rapid for individuals of low trait levels, due largely to their sub-optimal proportion of correct responses, resulting in shorter mean test length;

3. The shorter the test length, the less opportunity the Bayesian estimation procedure has to converge to extreme trait level estimates;

4. Non-convergence combines with negative information in some correct responses to diminish severely the effectiveness of measurement in the low regions of the trait.

Some of the conclusions just stated are speculative. Specifically, neither proportion correct as a function of θ nor the differences $(b_g - \theta)$ were examined in this study. Both of these reflect the effectiveness of the tailoring process. McBride (1975), however, reported data which showed proportion correct to be monotonically related to θ in another simulation study of Owen's Bayesian strategy.

One goal of adaptive testing should be to achieve a constant high level of measurement effectiveness at all levels of θ. This objective is equivalent to a high, horizontal information function. The Study 4 results show that the Bayesian sequential testing strategy failed to achieve this goal despite an unrealistically favorable set of circumstances: the perfect item pool, error-free item parameters, and a scoring model perfectly congruent with the item response model. The shortcomings of the Bayesian trait estimate were attributed to the regression-like tendency of the sequential estimates themselves, which in turn results in inappropriate item selection for individuals whose trait levels are relatively high or low.

There are at least two methods of ameliorating this problem, both of which to some extent should lessen the bias of estimate at the extremes and improve the information properties of the trait estimates. The first method involves the assumption of a rectangular rather than a normal prior distribution of Θ. The second method would involve replacing the Bayesian item selection procedure with a mechanical (e.g., non-mathematical) branching procedure, which would be less sensitive to large errors in the current trait estimate in its choice of the next item to administer. Needless to say, both of these alternatives involve a considerable departure from Owen's elegant procedure.

Implications. In testing persons of any given ability level, an ideal adaptive testing strategy would select for administration the most informative items available at that level. If the item pool were adequate, the result would be that mean proportion correct would be approximately constant across ability levels, and the information curve of the ability estimates would be very high and almost flat. Such an adaptive test would make equally good discriminations at any level of the ability trait. It would also have approximately equivalent utility at any level at which discriminations were to be made. It is apparent from the foregoing discussion, especially from the data of Study 4, that the properties of the Bayesian sequential adaptive test fall somewhat short of this ideal. The research reported here has shown that the Bayesian procedure results in very high correlations of ability level and test scores but also results in ability estimates which are strongly biased in the extremes and which are maximally informative only in the middle region of ability. If a test user were concerned primarily with *ordering* examinees as to ability level, the Bayesian sequential adaptive procedure would seem quite satisfactory. However, the tendency of the Bayesian procedure to yield accurate measurement in the vicinity of the prior mean at the expense of relatively inferior measurement elsewhere, may mandate selecting an alternative adaptive strategy if the test user requires either equi-discriminating measurement over a wide ability range or accurate ability estimation for ability levels not near the mean. Simulation research by Vale & Weiss (1975) on the stradaptive ability test (Weiss, 1973) shows that adaptive testing strategy provides measurement with the desired characteristics. Other promising strategies for adaptive testing have been proposed by Lord (1975) and Samejima (1975).

References

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

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

Green, B.F. Discussion. In Proceedings of the Conference on Computerized Adaptive Testing. Washington, D.C. June, 1975.

Jensema, C.J. An application of latent trait mental test theory to the Washington Pre-College Testing Program. Unpublished doctoral dissertation. University of Washington, 1972.

Jensema, C.J. The validity of Bayesian tailored testing. Educational and Psychological Measurement, 1974, 34, 757-766.

Lord, F.M. Some test theory for tailored testing. In Holtzman, W.H. (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970 (Chapter 8).

Lord, F.M. The "ability" scale in item characteristic curve theory. Research Bulletin 74-19. Princeton, N.J.: Educational Testing Service, June, 1974.

Lord, F.M. A broad-range test of verbal ability. Research Bulletin 75-5. Princeton, N.J.: Educational Testing Service, 1975.

Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McBride, J.R. Problem: scoring adaptive tests. In Weiss, D.J. (Ed.) Computerized adaptive trait measurement: problems and prospects. Research Report 75-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1975.

Owen, R.J. A Bayesian approach to tailored testing. Research Bulletin 69-92. Princeton, N.J.: Educational Testing Service, 1969.

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

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-233.

Samejima, F. Behavior of the maximum likelihood estimate in a simulated tailored testing situation. Paper presented at the meeting of the Psychometric Society, Iowa City, April, 1975.

Urry, V.W. Individualized testing by Bayesian estimation. Research Bulletin 0171-177. Seattle: Bureau of Testing, University of Washington, 1971.

Urry, V.W. Computer-assisted testing: the calibration and evaluation of the verbal ability bank. Technical Study 74-3. Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974.

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

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

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

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation. University of Chicago, 1971.

APPENDIX: Supplementary Tables

Table A-1
Mean Test Scores ($\hat{\Theta}$), Mean Error of Estimate (e), and Mean Test Length (k) for Three Item Pool Configurations, at Each of 31 Trait Levels (Θ)

	Item Pool Configuration								
	r_{ab}+.71			r_{ab}.0			r_{ab}-.71		
Θ	$\hat{\Theta}$	e	k	$\hat{\Theta}$	e	k	$\hat{\Theta}$	e	k
-3.0	-2.39	.612	30	-2.47	.532	30	-2.30	.696	14
-2.8	-2.26	.545	30	-2.29	.513	30	-2.20	.601	14
-2.6	-2.06	.542	30	-2.25	.352	30	-2.17	.427	15
-2.4	-2.00	.404	30	-2.06	.342	30	-2.08	.317	15
-2.2	-1.81	.390	30	-1.94	.362	30	-1.93	.269	16
-2.0	-1.70	.296	30	-1.80	.204	30	-1.74	.263	17
-1.8	-1.60	.200	30	-1.66	.141	30	-1.65	.146	18
-1.6	-1.44	.163	30	-1.45	.151	30	-1.48	.125	18
-1.4	-1.24	.162	30	-1.32	.082	30	-1.29	.110	20
-1.2	-1.12	.076	30	-1.12	.082	30	-1.14	.060	21
-1.0	-.93	.073	30	-.93	.071	30	-.98	.018	22
-.8	-.74	.055	30	-.74	.055	30	-.76	.037	24
-.6	-.56	.038	30	-.59	.014	30	-.58	.015	26
-.4	-.44	-.040	30	-.40	.004	30	-.35	.049	27
-.2	-.25	-.046	30	-.21	-.010	30	-.14	.062	29
0	-.06	-.058	30	.05	.046	30	.02	.021	30
.2	.20	-.003	30	.16	-.039	30	.19	-.007	30
.4	.35	-.053	30	.34	-.056	30	.35	-.051	30
.6	.53	-.068	29	.61	.010	30	.58	-.015	30
.8	.76	-.044	29	.74	-.058	30	.81	.013	30
1.0	.95	-.051	28	.89	-.113	30	.92	-.080	30
1.2	1.11	-.091	27	1.16	-.036	30	1.15	-.047	30
1.4	1.37	-.034	26	1.33	-.068	30	1.25	-.150	30
1.6	1.53	-.074	26	1.48	-.117	30	1.46	-.140	30
1.8	1.73	-.070	25	1.68	-.123	30	1.64	-.165	30
2.0	1.89	-.113	24	1.88	-.119	30	1.78	-.224	30
2.2	2.09	-.107	24	2.05	-.146	30	1.98	-.224	30
2.4	2.27	-.132	23	2.22	-.176	30	2.13	-.270	30
2.6	2.47	-.126	23	2.37	-.230	30	2.33	-.273	30
2.8	2.63	-.168	23	2.57	-.230	30	2.43	-.273	30
3.0	2.81	-.189	23	2.72	-.282	30	2.57	-.426	30

Table A-2

Estimated Value of the Derivative $\frac{\partial\hat{\theta}}{\partial\theta}$, Conditional Standard Deviation $\sigma_{\hat{\theta}|\theta}$ and Value of the Information Function $I_{\hat{\theta}}(\theta)$ for Three Item Pool Configurations, at Each of 31 Trait Levels (θ)

	Item Pool Configuration								
	r_{ab} +.71			r_{ab} 0			r_{ab} -.71		
θ	$\frac{\partial\hat{\theta}}{\partial\theta}$	$\sigma_{\hat{\theta}\|\theta}$	$I_{\hat{\theta}}(\theta)$	$\frac{\partial\hat{\theta}}{\partial\theta}$	$\sigma_{\hat{\theta}\|\theta}$	$I_{\hat{\theta}}(\theta)$	$\frac{\partial\hat{\theta}}{\partial\theta}$	$\sigma_{\hat{\theta}\|\theta}$	$I_{\hat{\theta}}(\theta)$
-3.0	.523	.307	2.90	.588	.336	2.58	.450	.353	1.63
-2.8	.566	.353	2.57	.629	.333	3.57	.511	.308	2.75
-2.6	.607	.328	3.42	.668	.304	4.83	.568	.279	4.14
-2.4	.645	.341	3.58	.704	.283	6.20	.621	.264	5.54
-2.2	.682	.321	4.51	.738	.294	6.31	.670	.268	6.26
-2.0	.716	.330	4.71	.770	.284	7.35	.716	.289	6.14
-1.8	.748	.324	5.33	.799	.228	12.29	.758	.289	6.87
-1.6	.778	.257	6.26	.826	.266	9.64	.796	.247	10.37
-1.4	.783	.311	6.34	.850	.265	10.29	.830	.230	13.01
-1.2	.832	.314	7.01	.872	.261	11.16	.860	.251	11.73
-1.0	.855	.278	9.46	.892	.275	10.52	.886	.235	14.21
- .8	.876	.316	7.69	.909	.278	10.70	.908	.244	13.86
- .6	.895	.283	10.00	.924	.260	12.63	.927	.244	14.44
- .4	.912	.282	10.47	.936	.288	10.57	.9[illegible]2	.255	1[illegible].66
- .2	.927	.308	9.06	.946	.278	11.59	.953	.284	1[illegible].96
0	.940	.305	9.50	.954	.249	14.68	.960	.257	1[illegible].96
.2	.946	.253	13.98	.959	.248	14.96	.963	.284	11.50
.4	.959	.255	14.14	.962	.281	11.72	.963	.252	14.59
.6	.965	.287	11.29	.962	.275	12.25	.958	.285	11.31
.8	.965	.269	12.86	.960	.248	15.00	.950	.276	11.85
1.0	.971	.228	18.15	.956	.250	14.62	.938	.336	7.79
1.2	.971	.228	18.13	.949	.250	14.42	.922	.294	9.84
1.4	.968	.218	19.71	.940	.272	11.94	.902	.295	9.36
1.6	.964	.246	15.35	.928	.259	12.85	.879	.301	8.52
1.8	.957	.229	17.46	.914	.292	9.81	.851	.317	7.21
2.0	.948	.263	13.00	.898	.289	9.66	.820	.296	7.67
2.2	.937	.230	16.56	.879	.260	11.43	.785	.321	5.98
2.4	.924	.210	19.35	.858	.255	11.32	.746	.294	6.44
2.6	.908	.227	16.00	.834	.270	9.55	.703	.349	4.06
2.8	.891	.258	16.69	.808	.250	10.46	.657	.332	3.91
3.0	.871	.218	16.00	.780	.279	7.82	.606	.293	4.28

DISTRIBUTION LIST

Navy

4 Dr. Marshall J. Farr, Director
Personnel and Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: Dr. James Lester

1 ONR Branch Office
1030 East Green Street
Pasadena, CA 91101
ATTN: Dr. Eugene Gloye

1 ONR Branch Office
536 South Clark Street
Chicago, IL 60605
ATTN: Dr. Charles E. Davis

1 Dr. M. A. Bertin, Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco 96503

1 Dr. H. Wallace Sinaiko
c/o Office of Naval Research
Code 450
Arlington, VA 22217

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Technical Director
Navy Personnel Research
and Development Center
San Diego, CA 92152

1 Assistant for Research Liaison
Bureau of Naval Personnel (Pers Or)
Room 1416, Arlington Annex
Washingtoh, DC 20370

1 Assistant Deputy Chief of Naval
Personnel for Retention Analysis
and Coordination (Pers 12)
Room 2403, Arlington Annex
Washington, DC 20370

1 LCDR Charles J. Theisen, Jr., MSC, USN
4024
Naval Air Development Center
Warminster, PA 18974

1 Dr. Lee Miller
Naval Air Systems Command
AIR-413E
Washington, DC 20361

1 CDR Paul D. Nelson, MSC, USN
Naval Medical R&D Command (Code 44)
National Naval Medical Center
Bethesda, MD 20014

1 Commanding Officer
Naval Health Research Center
San Diego, CA 92152
ATTN: Library

1 Chairman
Behavioral Science Department
Naval Command & Management Division
U.S. Naval Academy
Annapolis, MD 21402

1 Chief of Naval Education & Training
Naval Air Station
Pensacola, FL 32508
ATTN: CAPT Bruce Stone, USN

1 Mr. Arnold I. Rubinstein
Human Resources Program Manager
Naval Material Command (0344)
Room 1044, Crystal Plaza #5
Washington, DC 20360

1 Dr. Jack R. Borsting
U. S. Naval Postgraduate School
Department of Operations Research
Monterey, CA 93940

1 Director, Navy Occupational Task
Analysis Program (NOTAP)
Navy Personnel Program Support
Activity
Building 1304, Bolling AFB
Washington, DC 20336

1 Office of Civilian Manpower Management
Code 64
Washington, DC 20390
ATTN: Dr. Richard J. Niehaus

1 Office of Civilian Manpower Management
Code 263
Washington, DC 20390

1 Chief of Naval Reserve
Code 3055
New Orleans, LA 70146

1 Assistant to the Assistant Deputy Chief
of Naval Operations (Manpower)
Head, NAMPS Project Office
Room 1606, Arlington Annex
Washington, DC 20370
ATTN: Dr. Harry M. West

1 Superintendent
Naval Postgraduate School
Monterey, CA 93940
ATTN: Library (Code 2124)

1 Commander, Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203
ATTN: Code 015

1 Mr. George N. Graine
Naval Sea Systems Command
SEA 047C12
Washington, DC 20362

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. Norman J. Kerr

1 Principal Civilian Advisor
for Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508
ATTN: Dr. William L. Maloy

1 Director
Training Analysis & Evaluation Group
Code N-00t
Department of the Navy
Orlando, FL 32813
ATTN: Dr. Alfred F. Smode

1 Chief of Naval Training Support
Code N-21
Building 45
Naval Air Station
Pensacola, FL 32508

1 Naval Undersea Center
Code 6032
San Diego, CA 92132
ATTN: W. Gary Thomson

1 LCDR C. F. Logan, USN
F-14 Management System
COMFITAEWWINGPAC
NAS Miramar, CA 92145

1 Navy Personnel Research
and Development Center
Code 01
San Diego, CA 92152

5 Navy Personnel Research
and Development Center
Code 02
San Diego, CA 92152
ATTN: A. A. Sjoholm

2 Navy Personnel Research
and Development Center
Code 304
San Diego, CA 92152
ATTN: Dr. John Ford

2 Navy Personnel Research
and Development Center
Code 310
San Diego, CA 92152
ATTN: Dr. Martin F. Wiskoff

2 Navy Personnel Research
and Development Center
Code 309
San Diego, CA 92152
ATTN: Dr. C. Cory

1 Navy Personnel Research
and Development Center
San Diego, CA 92152
ATTN: Library

1 Navy Personnel Research
and Development Center
Code 9041
San Diego, CA 92152
ATTN: Dr. J. D. Fletcher

1 D. M. Gragg, CAPT, MC, USN
Head, Educational Programs Development
Department
Naval Health Sciences Education and
Training Command
Bethesda, MD 20014

Army

1 Technical Director
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Headquarters
U.S. Army Administration Center
Personnel Administration Combat
Development Activity
ATCP-HRQ
Ft. Benjamin Harrison, IN 46249

1 Armed Forces Staff College
Norfolk, VA 23511
ATTN: Library

1 Commandant
U.S. Army Infantry School
Fort Benning, GA 31905
ATTN: ATSH-DET

1 Deputy Commander
U.S. Army Institute of Administration
Fort Benjamin Harrison, IN 46216
ATTN: EA

1 Dr. Stanley L. Cohen
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph Dusek
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Leon H. Nawrocki
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Joseph Ward
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 HQ USAREUR & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

1 ARI Field Unit - Leavenworth
Post Office Box 3122
Fort Leavenworth, KS 66027

1 Mr. James Baker
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. James L. Raney
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton S. Katz, Chief
Individual Training & Performance
Evaluation
U.S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Air Force

1 Research Branch
AF/DPMYAR
Randolph AFB, TX 78148

1 Dr. G. A. Eckstrand (AFHRL/AST)
Wright-Patterson AFB
Ohio 45433

1 AFHRL/DOJN
Stop #63
Lackland AFB, TX 78236

1 Dr. Martin Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Dr. Alfred R. Fregly
AFOSR/NL
1400 Wilson Boulevard
Arlington, VA 22209

1 AFHRL/PED
Stop #63
Lackland AFB, TX 78236

1 Major Wayne S. Sellman
Chief of Personnel Testing
HQ USAF/DPMYP
Randolph AFB, TX 78148

Marine Corps

1 Director, Office of Manpower Utilization
Headquarters, Marine Corps (Code MPU)
MCB (Building 2009)
Quantico, VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code RD-1)
Headquarters, U.S. Marine Corps
Washington, DC 20380

1 Chief, Academic Department
Education Center
Marine Corps Development and Education Command
Marine Corps Base
Quantico, VA 22134

1 Mr. E. A. Dover
2711 South Veitch Street
Arlington, VA 22206

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-1/62)
U.S. Coast Guard Headquarters
Washington, DC 20590

Other DOD

1 Military Assistant for Human Resources
Office of the Secretary of Defense
Room 3D129, Pentagon
Washington, DC 20301

1 Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Human Resources Research Office
1400 Wilson Boulevard
Arlington, VA 22209

1 Dr. Robert Young
Advanced Research Projects Agency
Human Resources Research Office
1400 Wilson Boulevard
Arlington, VA 22209

1 Mr. Frederick W. Suffa
Chief, Recruiting and Retention Evaluation
Office of the Assistant Secretary of Defense, M&RA)
Room 3D970, Pentagon
Washington, DC 20301

12 Defense Documentation Center
Cameron Station, Building 5
Alexandria, VA 22314
ATTN: TC

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E Street, N.W.
Washington, DC 20415

1 Dr. William Gorham, Director
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E Street, N.W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E Street, N.W.
Washington, DC 20415

1 Dr. Harold T. Yahr
Personnel Research and Development Center
U.S. Civil Service Commission
1900 E Street, N.W.
Washington, DC 20415

1 Dr. Richard C. Atkinson
Deputy Director
National Science Foundation
1800 G Street, N.W.
Washington, DC 20550

1 Dr. Andrew R. Molnar
Technological Innovations in Education Group
National Science Foundation
1800 G Street, N.W.
Washington, DC 20550

1 U.S. Civil Service Commission
Federal Office Building
Chicago Regional Staff Division
Regional Psychologist
230 South Dearborn Street
Chicago, IL 60604
ATTN: C. S. Winiewicz

1 Dr. Carl Frederiksen
Learning Division, Basic Skills Group
National Institute of Education
1200 19th Street, N.W.
Washington, DC 20208

Miscellaneous

1 Dr. Scarvia B. Anderson
Educational Testing Service
17 Executive Park Drive, N.E.
Atlanta, GA 30329

1 Dr. John Annett
Department of Psychology
The University of Warwick
Coventry CV47AL
ENGLAND

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

1 Dr. Barnard M. Bass
University of Rochester
Graduate School of Management
Rochester, NY 14627

1 Century Research Corporation
4113 Lee Highway
Arlington, VA 22207

1 Dr. Kenneth E. Clark
University of Rochester
College of Arts and Sciences
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
University of Southern California
Department of Psychology
University Park
Los Angeles, CA 90007

1 Dr. Allan M. Collins
Bolt Beranek and Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

1 Dr. Rene' V. Dawis
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Ruth Day
Yale University
Department of Psychology
2 Hillhouse Avenue
New Haven, CT 06520

1 Dr. Norman R. Dixon
200 South Craig Street
University of Pittsburgh
Pittsburgh, PA 15260

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 ERIC
Processing and Reference Facility
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Victor Fields
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. Edwin A. Fleishman
Visiting Professor
University of California
Graduate School of Administration
Irvine, CA 92664

1 Dr. Robert Glaser, Co-Director
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15213

1 Mr. Harry H. Harman
Educational Testing Service
Princeton, NJ 08540

1 Dr. Richard S. Hatch
Decision Systems Associates, Inc.
5640 Nicholson Lane
Rockville, MD 20852

1 Dr. M. D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO Central Division
400 Plaza Building
Pace Boulevard at Fairfield Drive
Pensacola, FL 32505

1 HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921
ATTN: Library

1 HumRRO Central Division/Columbus Office
Suite 23, 2601 Cross Country Drive
Columbus, GA 31906

1 HumRRO
Joseph A. Austin Building
1939 Goldsmith Lane
Louisville, KY 40218

1 Dr. Lawrence B. Johnson
Lawrence Johnson & Associates, Inc.
2001 S Street, N.W., Suite 502
Washington, DC 20009

1 Dr. Roger A. Kaufman
203 Dodd Hall
Florida State University
Tallahassee, FL 32306

1 Dr. David Klahr
Carnegie-Mellon University
Department of Psychology
Pittsburgh, PA 15213

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Corton Drive
Santa Barbara Research Park
Goleta, CA 93017

1 Dr. William C. Mann
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marine Del Rey, CA 90291

1 Mr. Edmond Marks
315 Old Main
Pennsylvania State University
University Park, PA 16802

1 Dr. Leo Munday, Vice President
American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Donald A. Norman
University of California, San Diego
Department of Psychology
LaJolla, CA 92037

1 Mr. A. J. Pesch, President
Eclectech Associates, Inc.
P.O. Box 178
North Stonington, CT 06359

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, CA 90265

Dr. Joseph W. Rigney
University of Southern California
Behavioral Technology Laboratories
3717 South Grand
Los Angeles, CA 90007

1 Dr. George E. Rowland
Rowland and Company, Inc.
P.O. Box 61
Haddonfield, NJ 08033

1 Dr. Benjamin Schneider
University of Maryland
Department of Psychology
College Park, MD 20742

1 Dr. Arthur I. Siegel
Applied Psychological Services
404 East Lancaster Avenue
Wayne, PA 19087

1 Dr. Henry P. Sims, Jr.
Room 630 - Business
Indiana University
Bloomington, IN 47401

1 Dr. Richard Snow
Stanford University
School of Education
Stanford, CA 94305

1 Mr. George Wheaton
American Institutes for Research
3301 New Mexico Avenue, N.W.
Washington, DC 20016

1 Dr. K. Wescourt
Stanford University
Institute for Mathematical Studies
in the Social Sciences
Stanford CA 94305

1 Richard T. Mowday
College of Business Administration
University of Nebraska, Lincoln
Lincoln, NE 68588

1 Dr. John J. Collins
Vice President
Essex Corporation
6305 Caminito Estrellado
San Diego, CA 92120

1 Dr. Lyle Schoenfeldt
Department of Psychology
University of Georgia
Athens, Georgia 30602

1 Dr. Patrick Suppes, Director
Institute for Mathematical Studies
in the Social Sciences
Stanford University
Stanford, CA 94305

Previous Reports in this Series

73-1. Weiss, D.J. & Betz, N.E. Ability Measurement: Conventional or Adaptive? February 1973. (AD 757788).

73-2. Bejar, I.I. & Weiss, D.J. Comparison of Four Empirical Differential Item Scoring Procedures. August 1973.

73-3. Weiss, D.J. The Stratified Adaptive Computerized Ability Test. September 1973. (AD 768376).

73-4. Betz, N.E. & Weiss, D.J. An Empirical Study of Computer-Administered Two-Stage Ability Testing. October 1973. (AD 768993).

74-1. DeWitt, L.J. & Weiss, D.J. A Computer Software System for Adaptive Ability Measurement. January 1974. (AD 773961).

74-2. McBride, J.R. & Weiss, D.J. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974. (AD 781894).

74-3. Larkin, K.C. & Weiss, D.J. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974. (AD 783553).

74-4. Betz, N.E. & Weiss, D.J. Simulation Studies of Two-Stage Ability Testing. October 1974. (AD A001230).

74-5. Weiss, D.J. Strategies of Adaptive Ability Measurement. December 1974. (AD A004270).

75-1. Larkin, K.C. & Weiss, D.J. An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing. February 1975. (AD A006733).

75-2. McBride, J.R. & Weiss, D.J. TETREST: A FORTRAN IV Program for Calculating Tetrachoric Correlations. March 1975. (AD A007572).

75-3. Betz, N.E. & Weiss, D.J. Empirical and Simulation Studies of Flexilevel Ability Testing. July 1975. (AD A013185).

75-4. Vale, C.D. & Weiss, D.J. A Study of Computer-Administered Stradaptive Ability Testing. October 1975. (AD A018758).

75-5. Weiss, D.J. (Ed.). Computerized Adaptive Trait Measurement: Problems and Prospects. November 1975. (AD A018675).

75-6. Vale, C.D. & Weiss, D.J. A Simulation Study of Stradaptive Ability Testing. December 1975. (AD A020961).

AD Numbers are those assigned by the Defense Documentation Center, for retrieval through the National Technical Information Service.

Copies of these reports are available, while supplies last, from:

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455