Some Properties of a Bayesian Adaptive Ability Testing Strategy

James R. McBride
University of Minnesota

Four monte carlo simulation studies of Owen's Bayesian sequential procedure for adaptive mental testing were conducted. In contrast to previous simulation studies of this procedure which 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. The properties of interest included the conditional bias of the ability 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 testing 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.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 1 Winter 1977 pp. 121–140

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, 1976).

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 among available items 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^2, 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. Jensema (1976) presents equations used for ability estimation and item selection in Owen's procedure.

The test score resulting from administration of a test based on Owen's procedure is a statistical estimation of the examinee's latent ability. Owen (1975) showed that estimator to be a consistent one; i.e., as the number of items administered approaches infinity, the value of the ability estimate approaches the true ability. Practically speaking, of course, the number of items administered never will approach infinity; but if the pool of available items is sufficiently large and appropriately constituted, σ_m^2 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, 1976) monte carlo simulation studies of Owen's Bayesian testing strategy have evaluated its merit solely in terms of the "fidelity" (or "validity")[1] 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.

Still another criterion for evaluating adaptive tests is the information function (Birnbaum, 1968). The information in a set of test scores (x) can be defined as

$$I_x(\Theta) = \left[\frac{\frac{\partial}{\partial\Theta}(E(x|\Theta))}{\sigma_{x|\Theta}}\right]^2 . \qquad [1]$$

[1]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.

The numerator in Equation 1 is the first partial derivative of the function describing the regression of test scores (x) on the trait (Θ). The denominator in Equation 1 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 (Θ) are known.

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.

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 information 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 the bias and information 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.

Method

Examinees. For the purposes of monte carlo simulation, an "examinee" i is characterized by a numerical value, which is his actual ability level, Θ_i. In each of the studies below examinees were simulated by specifying a set of values Θ_i.

Test items. For each separate item administration an artificial item g was simulated. In each of the studies, items were simulated by specifying their parameters a_g, b_g and c_g (i.e., their discrimination, difficulty and guessing parameters, respectively).

For Study 1, two 100-item "ideal" item pools were simulated, corresponding to the ones used by Jensema (1972). For Studies 2, 3 and 4, however, an infinite item pool was simulated. In each of the last three studies the discrimination (a) and guessing (c) parameters were held constant, and a hypothetical item was generated whose difficulty was equal to:

$$b_g = M - \frac{1}{1.7\alpha_g} \log \left[\frac{1+(1+8c_g)^{1/2}}{2} \right], \quad [2]$$

In Equation 2, M is the current Bayesian ability estimate. The formula gives the item difficulty value having maximal information value when $\Theta_i=M$, given the values of a_g and c_g (Birnbaum, 1968, p. 464).

Item responses. The dichotomous (0,1) score of any examinee on any item is a probabilistic function of his ability status Θ_i, the item difficulty b_g, and the parameters a_g and c_g. The probability $P_g(\Theta_i)$ of a correct response under the logistic model item characteristic curve is

$$P_g(\Theta_i) = c_g + (1-c_g) / (1+ \exp(-1.7a_g(\Theta_i-b_g))). \quad [3]$$

This probability differs from that of the normal ogive response model by less than .01 at any point.

In order to simulate dichotomous item responses, each time an item was administered, the quantity $P_g(\Theta_i)$ was compared with a pseudo-random number R_{gi} sampled from a distribution uniform in the interval [0,1]. An item score of 1 ("correct") was assigned if $P_g(\Theta_i)$ equalled or exceeded R_{gi}; otherwise a score of 0 was assigned.

Study 1: An Ideal Item Pool with Variable Test Length

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 $-2.4 \leqslant b \leqslant 2.4$; 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_m^2 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). 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

Varibles. Dependent variables were the individual ability estimates ($\hat{\Theta}$) and the number of items (k) required to satisfy the posterior variance termination criterion of $\sigma_m^2 \leqslant .0625$. 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 $a=1.60$.

Procedure. 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 para-

meters (μ_m, σ_m^2) 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_m^2 was smaller than .0625, whichever occurred first. Once any particular item had been taken by the examinee it was not reused.

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;

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\Theta}_i-\Theta_i) \qquad [4]$$

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

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{\Theta}_i-\Theta_i| \qquad [5]$$

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. $r_{\hat{\theta}\theta}$, the fidelity coefficient;
f. the mean, minimum and maximum test length required to achieve the posterior variance termination criterion.

Results

Table 1 contains the results from Study 1. As Table 1 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 1
Properties of the Bayesian Sequential Test for Two Values of Item Discrimination, with Corrected Guessing and Ideal Item Pool

	Item Discrimination (a)	
Property	.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. The differences likely are 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

The discovery in Study 1 of positive bias in the Bayesian trait estimates, 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 an infinite 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 two different guessing conditions described below—no guessing and corrected guessing. Under each of these conditions fourteen infinite 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 examined in Study 1 were of interest in Study 2.

Method

Dependent variables in Study 2 were the same as in Study 1. Independent variables were the discriminating power of the item pool and the effect of guessing. To study the effect of guessing, two 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; 2) 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".

"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. 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.

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 ($\sigma^2_m \leqslant .0625$) 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.

A total of 28 test administration conditions was simulated—14 "item pools" under each of the two guessing conditions. For each test administration, the same sequential test properties estimated in Study 1 were estimated.

Results

No-guessing condition. As Table 2 shows, test length was constant within item discrimination level under no-guesing, and diminished inversely with level of item discrimination. The posterior variance termination criterion was reached for 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 $a \geqslant 1.00$. 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 2). 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 pools to yield *larger* average errors than the moderately discriminating item pools.

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.

Guessing condition. As Table 3 shows, 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.

Table 3 also shows that 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.

As was seen in Study 1, test length correlated

Table 2

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.

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 Corrected 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	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.

strongly with ability (and ability estimates) where it was free to vary (Table 3). 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.

As in the no-guessing condition, Table 3 also shows that errors of estimate correlated negatively (−.25 to −.42) with ability and with ability estimates (−.09 to −.23). Similarly, 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 twenty-eight bias estimates was negative.

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, 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

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 2 and 3). 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 2 and 3). 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 infinite item pool again was employed.

Method

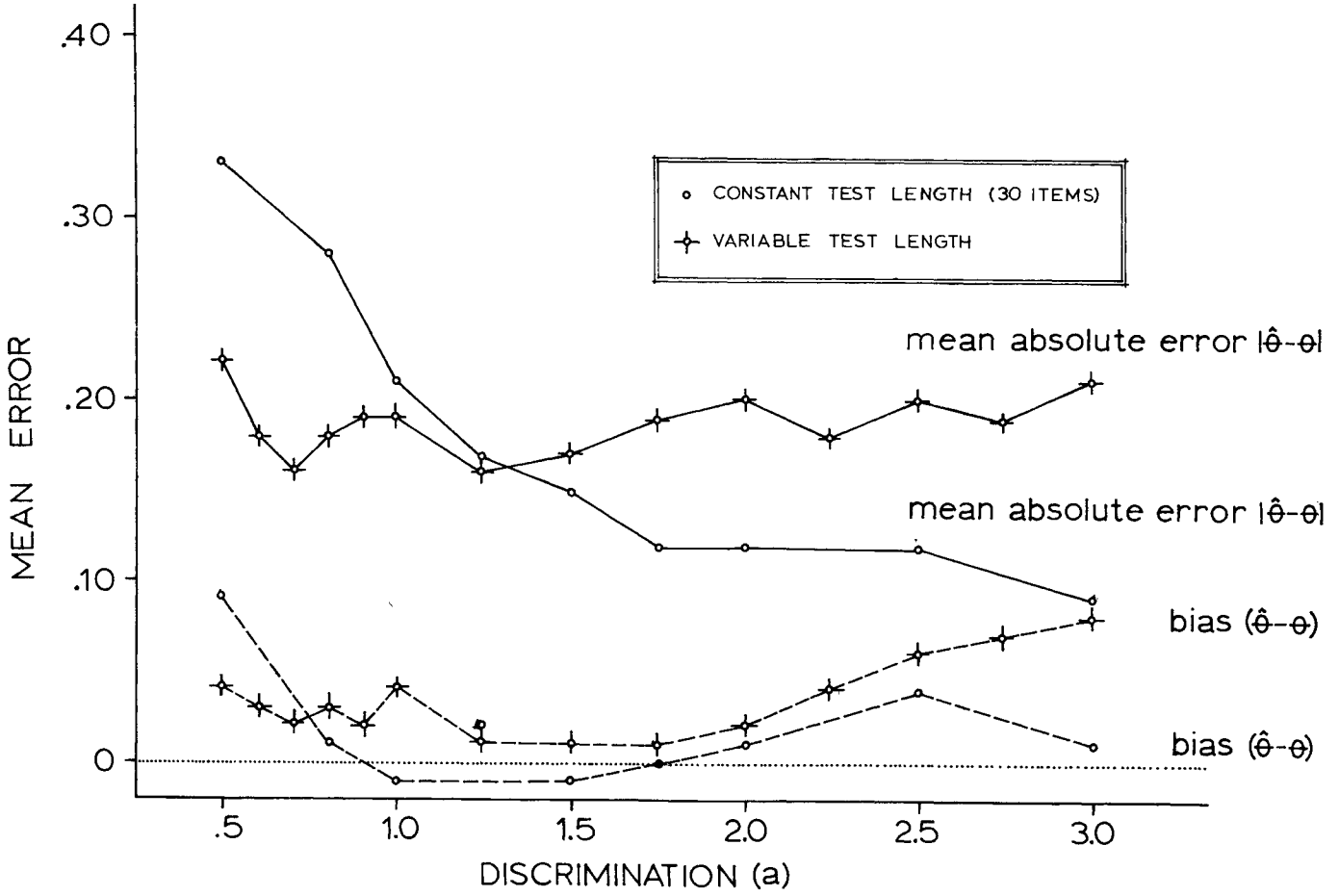
Nine levels of discriminating power were studied: a_g=.6, .8, 1.0, 1.25, 1.50, 1.75, 2.0, 2.5, 3.0. Examinees were the same 100 simulated ability values used in Studies 1 and 2. Infinite item pools were simulated, as described in Study 2.

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. The same test properties described in Study 1 were estimated. Additionally, the correlations of the posterior variance with Θ and $\hat{\Theta}$ were calculated.

Results

Table 4 and Figure 1 contain the results of

Figure 1
Mean Absolute Error and Bias for Two Different Test Termination Criteria



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 1.

As Figure 1 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 $(\Theta_i - \hat{\Theta}_i)$ correlated negatively with Θ and $\hat{\Theta}$, suggesting a regression effect.

As Table 4 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=.86 for a=.6, and r=.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. The fidelity coefficient increased with the item discriminating power, from .93 at a=.60 to .99 at a=1.5 and higher.

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 discrimination 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

This study examined properties of the Bayesian sequential testing strategy as a function of ability level. These properties include the conditional bias of the ability estimates, mean test length, and the "information" (Birnbaum, 1968) in the Bayesian test ability estimates.

This study also 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 pre-

Table 4
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

	Item Discrimination (a)								
Property	.6	.8	1.0	1.25	1.5	1.75	2.0	2.5	3.0
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

vious 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 item pool configurations. Each 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

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 $[-3.0 \leqslant \Theta \leqslant +3.0]$. 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 infinite item pools were simulated, one for each configuration. The three configurations studied included one with a moderate (r=.71) positive correlation of a with b (referred to hereinafter as $r_{ab}+$), one with a negative correlation ($r_{ab}-$) of the same magnitude, 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 ($\bar{a}$), which was 1.25 for all configurations. In all cases, c was .20. Details of the simulation of item pool configuration are given by McBride & Weiss (1976, pp. 20-21).

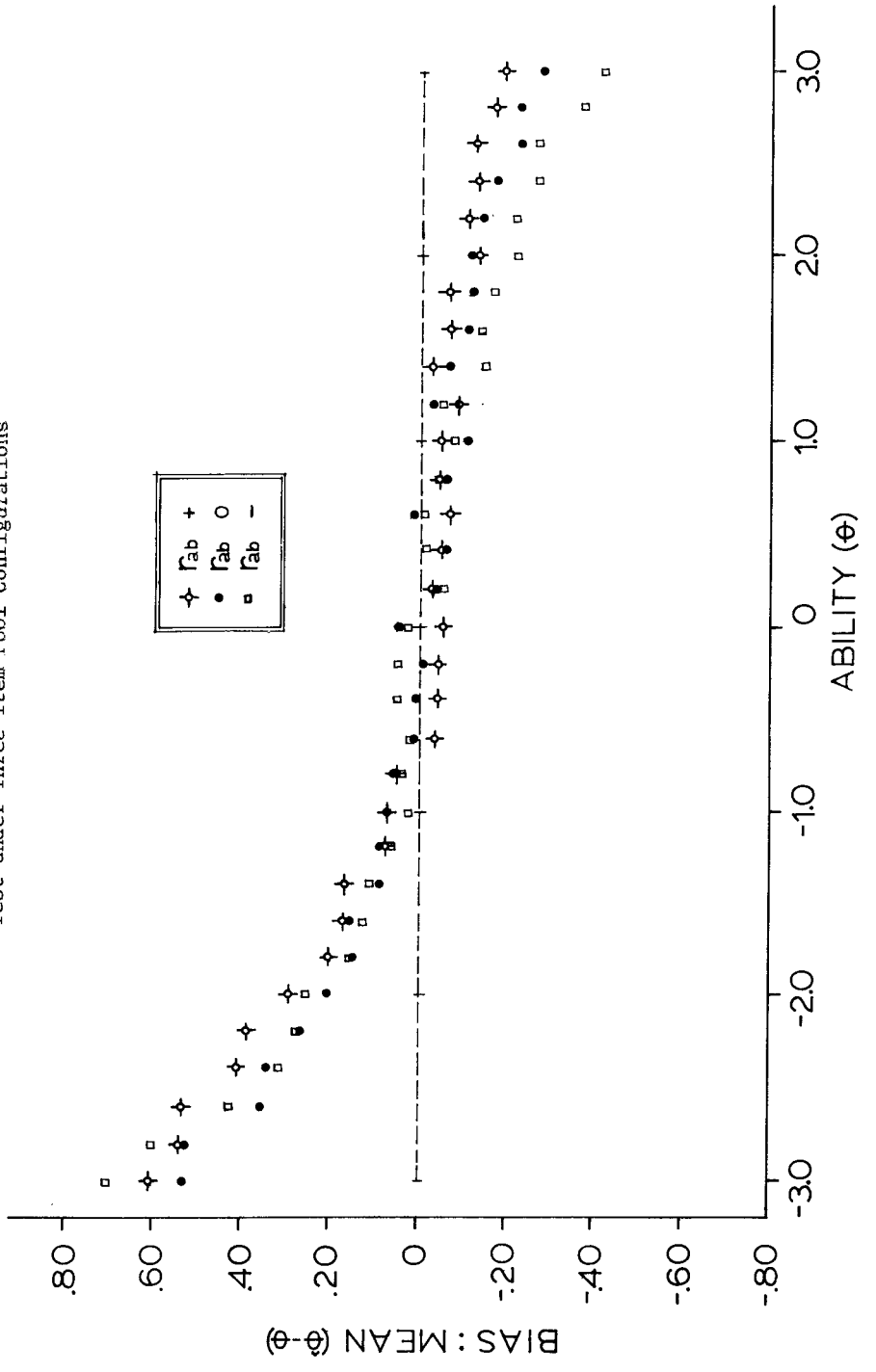
Item responses and test administration were simulated in the same manner described in Study 1. A posterior variance termination criterion of $\sigma^2_m \leqslant .0625$ was used, with an arbitrary maximum test length of 30 items. The corrected-guessing condition was used.

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,

$$\bar{\hat{\Theta}}_i \mid \Theta_i = \frac{1}{100} \Sigma \hat{\Theta}_i \qquad [6]$$

Figure 2
Mean Error of Estimate ($\hat{\theta}-\theta$) at 31 Ability Points ($\hat{\theta}$) for the Simulated Bayesian Sequential Test under Three Item Pool Configurations



b. the conditional variance,

$$\sigma^2_{\hat{\Theta}_i|\Theta_i} = \frac{1}{100}\Sigma(\hat{\Theta}_i - \bar{\hat{\Theta}}_i)^2 \quad [7]$$

c. the conditional bias,

$$b_{i|\Theta_i} = \bar{\hat{\Theta}}_i - \Theta_i \quad [8]$$

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

In the calculation of values of information using Equation 1, 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 denominator of Equation 1 was estimated from Equation 7.

Results

Bias. Figure 2 contains the plot of conditional bias on ability (numerical values are given by McBride & Weiss, 1976, p. 33). For each configuration, the curve described by these data is non-linear. The conditional bias for all three configurations was close to zero for $-1 \leqslant \Theta \leqslant 1$, 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 affected systematically by the configuration of the item pool.

Table 5 contains plots of mean test length as a function of ability level for each item pool configuration. For the $r_{ab}0$ configuration, test length was constant at 30 items, the arbitrary maximum. For $r_{ab}+$, where the most discriminating items were available at the higher difficulty

Table 5
Mean Number of Items to Termination (Test Length) at 31 Ability Points (Θ) for the Simulated Bayesian Sequential Test under Three Item Pool Configurations

	Item Pool Configuration		
Ability (Θ)	$r_{ab}+$	$r_{ab}0$	$r_{ab}-$
-3.0	30	30	14
-2.8	30	30	14
-2.6	30	30	15
-2.4	30	30	15
-2.2	30	30	16
-2.0	30	30	17
-1.8	30	30	18
-1.6	30	30	18
-1.4	30	30	20
-1.2	30	30	21
-1.0	30	30	22
- .8	30	30	24
- .6	30	30	26
- .4	30	30	27
- .2	30	30	29
0	30	30	30
.2	30	30	30
.4	30	30	30
.6	29	30	30
.8	29	30	30
1.0	28	30	30
1.2	27	30	30
1.4	26	30	30
1.6	26	30	30
1.8	25	30	30
2.0	24	30	30
2.2	24	30	30
2.4	23	30	30
2.6	23	30	30
2.8	23	30	30
3.0	23	30	30

levels, test length was constant at 30 items for Θ levels less than .6, then declined gradually to a mean of 23 items at Θ=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.

Table 5 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 2) 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.

Information. Figure 3 contains smoothed information curves for the three item pool configurations. For the $r_{ab}0$ configuration the information curve shown in Figure 3 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 Θ=−3 is even lower, but it increases steadily—almost linearly—with Θ. The $r_{ab}+$ information curve surpasses that of $r_{ab}0$ at Θ≥+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 Θ=0. Surprisingly, the information curve thereafter decreases with Θ, reaching its lowest point at Θ=−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.

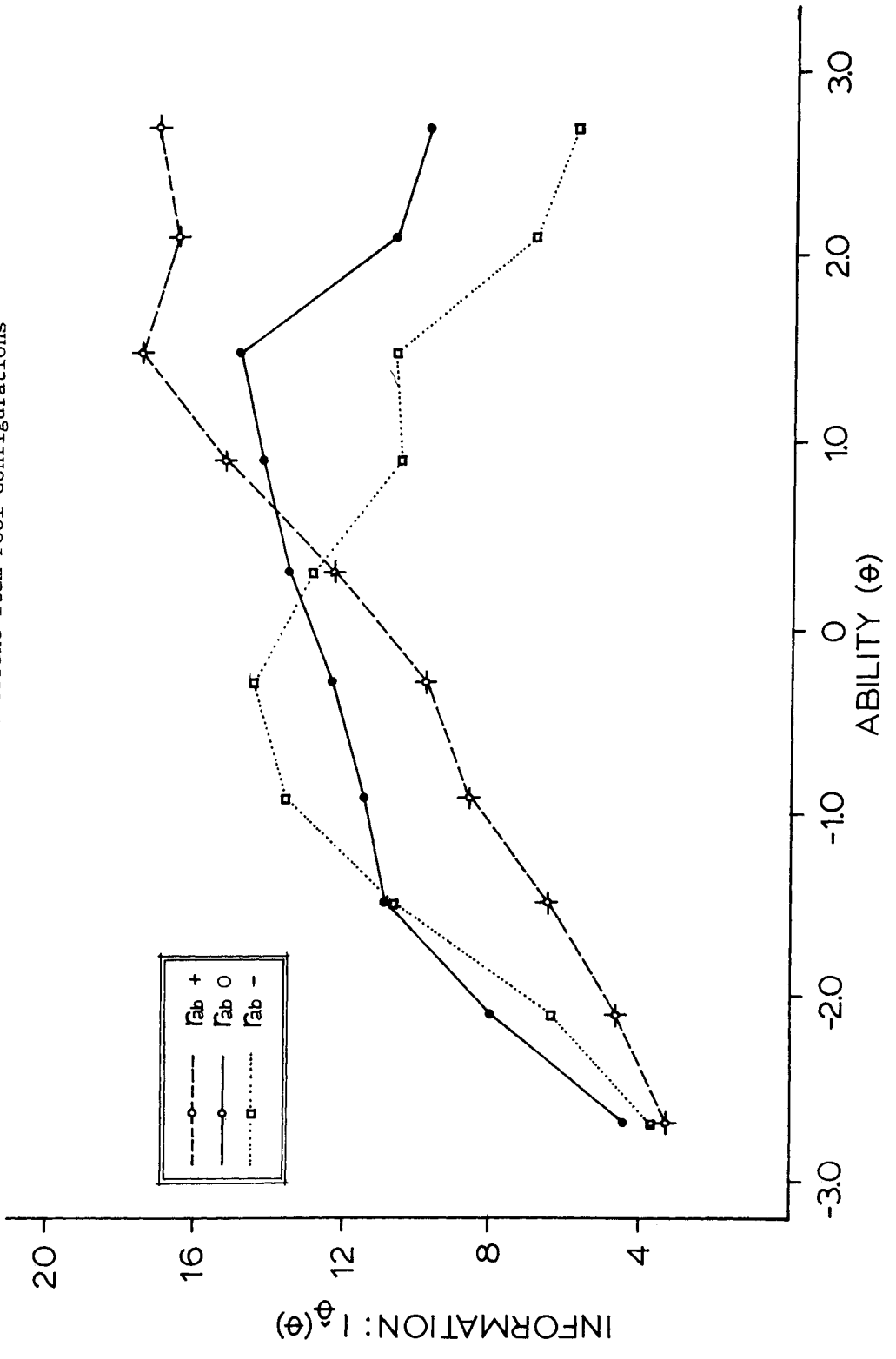
Summary and Conclusions

Previous research (e.g., Urry, 1971, 1974; Jensema, 1972, 1976) 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 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). When guessing was allowed in the model, bias was consistently positive, and increased as item discriminations increased and mean test length became very short.

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

Figure 3
Smoothed Information Curves for the Bayesian Sequential Test under Three Different Item Pool Configurations



In terms of fidelity coefficients, there was no appreciable difference between those obtained under no-guessing and under guessing conditions, given the common termination criterion.

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. 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. 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) 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 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. 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. Owen's 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 almost always will be too easy for extremely able examinees.

For low Θ the initial prior is an overestimate. Hence the first item selected generally will 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 Θ and thus result in another inappropriate choice of item difficulty. Furthermore, as Samejima (1973) has shown, when guessing is a factor there actually may 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.

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., nonmathematical) 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 (1976) 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 Psy-

chology, University of Minnesota, Minneapolis, 1974. (AD A018758).

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.: U.S. Civil Service Commission, March, 1976.

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.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement, 1,* 1976, 111-120.

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. *Applied Psychological Measurement, 1,* 1976, 95-99.

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

McBride, J. R. and Weiss, D. J. *Some properties of a Bayesian adaptive ability testing strategy.* Research Report 76-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1976. (AD A022964).

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, 70,* 1975, 351-356.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38,* 1973, 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. (AD A020961).

Weiss, D. J. *The stratified adaptive computerized ability test.* Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1973. (AD 768376).

Weiss, D. J. *Strategies of adaptive ability measurement.* Research Report 74-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1974. (AD A004270).

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

Acknowledgments

This research was supported by the Personnel and Training Programs, Office of Naval Research, under contract number N00014-76-C-0243, NR150-382, David J. Weiss, Principal Investigator. Portions of this paper were presented at the Spring 1975 meeting of the Psychometric Society, Iowa City, Iowa, April 24, 1975; and at the First Conference on Computerized Adaptive Testing, Washington, D.C., June 12, 1975.

Author's Address

James R. McBride, U.S. Army Research Institute, 1300 Wilson Boulevard, Arlington, VA 22209.