ADAPTIVE TESTING RESEARCH AT MINNESOTA — SOME PROPERTIES OF A BAYESIAN SEQUENTIAL ADAPTIVE MENTAL TESTING STRATEGY [1]

JAMES R. MCBRIDE
University of Minnesota

Adaptive or tailored testing subsumes a number of different strategies for adapting the difficulty of test items to the ability of the examinee. One of the most elegant of such strategies is a Bayesian sequential technique proposed by Owen (1969) and studied empirically by several investigators including Wood (1969), Urry (1971) and Jensema (1972).

Owen's technique is a general one for the sequential design and the 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 out of 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[2] (a_g, b_g, c_g) and estimation of ability via a Bayesian-motivated approximation. At each step m in the ability estimation sequence, a normal prior distribution on ability (θ) is assumed, with parameters ($\mu_m, \sigma^2{}_m$), 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 c_g=0 (i.e., no guessing) and 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 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, $\sigma^2{}_m$ of the prior distribution are updated in accord with the examinee's performance on the item. In the case of a correct answer:

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

and

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

Following a wrong answer

$$\mu_{m+1} = E(\theta|0) = \mu_m - \left(\frac{\sigma^2{}_m}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}}\right)\left(\frac{\phi(D)}{\Phi(D)}\right) \quad (2)$$

and

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

In the above equations (taken from Owen, 1975)

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

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

$$D = (b_g - \mu_m) \Big/ \sqrt{\frac{1}{a_g^2} + \sigma_g^2} \quad (3)$$

$$A = c_g + (1-c_g)\,\phi(-D)$$

[1] Research reported herein was supported by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under contract No. 00014-67-A-0113-0029, NR No. 150-343.

Portions of these results were presented at the Spring meeting of the Psychometric Society in Iowa City, Iowa, April 1975.

A complete report of these results is in preparation (McBride & Weiss, 1975a).

[2] As most commonly used, a_g and b_g respectively are the discrimination and difficulty parameters of the normal ogive model. C_g is the guessing parameter, the probability that an examinee will respond correctly to the item when he does not know the answer. The subscript g indexes items.

μ_{m+1} and $\sigma^2{}_{m+1}$, the parameters 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 taken to be normal, an assumption which is not strictly correct after the first item (Owen, 1975). Testing may be terminated when $\sigma^2{}_m$ 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 $\sigma^2{}_m$ is a measure of the uncertainty of the estimate. Urry (1971) and Jensema (1972, 1974) have interpreted $\sigma^2{}_m$ as the squared standard error of estimate (S.E.E.) of θ_i. Owen (1975) gives a theorem showing that as $m \to \infty$, $\mu_m \to \theta$.

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, $\sigma^2{}_m$ will diminish rapidly, permitting valid estimation of θ in a very 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 from 3 to 30 items the validity of a much longer conventional test, with the average number of items diminishing as their discriminatory power increased.

Validity, i.e., 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, in the same metric, of underlying trait level, the accuracy of the estimates is also an interesting datum. By "accuracy" here is meant the closeness of the estimates to actual ability, which 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. As a matter of convention, then, in the following the term "accuracy" will refer to mean absolute error of estimate, $(1/N)\ \Sigma'|\hat{\theta}_i-\theta_i|$; "bias" will refer to mean algebraic error of estimate $(1/N)\ \Sigma\ (\hat{\theta}_i-\theta_i)$; and "conditional bias" will refer to mean algebraic error of estimate at a given value of θ, $(1/N)\ \Sigma(\hat{\theta}_i-\theta|\theta)$.

The purpose of the present paper is to report the results of a series of simulation studies designed to investigate the influence of item pool characteristics on some properties of the Bayesian sequential test other than the correlational validity of the trait estimates. These properties will include bias and accuracy of the estimates, as well as others enumerated below.

The studies reported below were motivated by results obtained with live testing of Owen's strategy. Using a 329-item pool of vocabulary knowledge test items, a correlation of .80 was obtained between estimated ability and number of test items to termination (McBride & Weiss, 1975b). Simulation studies designed to investigate the influence of the item pool on that unexpectedly large correlation led to our discovery of systematic non-linear bias in the Bayesian estimates of ability. The nature of the bias, and some of its correlates, are discussed below.

METHOD

1. *Dependent variables* of interest included test length (number of test items administered before the termination criterion was reached), errors of estimate $(\hat{\theta}-\theta)$, bias of estimate (mean over individuals of $(\hat{\theta}-\theta)$), absolute value of the error $|\hat{\theta}-\theta|$, and validity of the estimates of θ, $r_{\theta\hat{\theta}}$.

2. *Independent variables* of interest included the effects of guessing in both the response model and the scoring algorithm, of item discrimination, and the correlation of difficulty and discrimination parameters in the item pool, and of different termination criteria.

3. *Examinees* for the first study were simulated by computer-generation of pseudorandom numbers (from a normal population with mean 0 and variance 1) which represented the ability θ_i of each examinee, i. For the second study, 100 examinees were simulated at each of 31 points on the ability continuum.

4. *Item responses* were simulated by comparing $P'_g(\theta_i)$ for each item g and examinee i with a random number e_{gi} from a rectangular distribution in the interval [0,1]. A score of 1 for examinee i on item g was assigned if $P'_g(\theta_i) \geq e_{gi}$. Otherwise a score of 0 was assigned.

5. *Item pools* were simulated under two different conditions:

a. A *perfect item pool* with items of constant discrimination a_g and guessing parameter c_g was simulated. Under this condition, the computer program computed the optimal difficulty b_{m+1} of the next item to administer, and a simulated item with that difficulty value was made available. This is referred to as a "perfect" item pool because in effect we have simulated an item pool in which an unlimited number of items is available at any point on the difficulty continuum. The estimated optimal difficulty of an item to administer at stage $m+1$ is equal to the current ability estimate, $\hat{\theta}_m$, when guessing is not a factor (i.e., when $c_g=0$). When guessing is a factor ($c_g>0$), the estimated optimal difficulty b_g is smaller than $\hat{\theta}_m$ by an amount which is a joint function of a_g and c_g. That is, when $c_g>0$ $(b_g-\hat{\theta}_m)<0$. (Actually, the true optimal difficulty is a function of a_g, c_g and the unknown parameter θ. The Bayesian sequential test procedure only estimates θ and hence estimates the optimal item difficulty. At any rate, the simulated "perfect" item pool makes available at every step m an item whose difficulty is exactly equal to the estimated optimal item difficulty based on a_g, c_g, and the then current estimate of θ).

b. A *differentially discriminating "perfect" item pool* was simulated by having unlimited item difficulties b_g available (as in a. above), but varying item discrimination systematically so that the mean a_g could be specified and

the regression of a_g of item difficulty b_g could be varied. In this way it was possible to simulate item pools in which more highly discriminating items were available in some regions of the ability continuum than in others. The details of this procedure are described in Study 2, below.

6. The Bayesian sequential test was simulated by a computer program. Input variables were θ_i; the parameters μ_0 and $\sigma^2{}_0$ of the initial prior distribution on θ; the number of items to be administered to any examinee; the constant discrimination parameter a_g of the *perfect item pool* (or the mean discrimination parameter of the *differentially discriminating perfect item pool)*, along with two guessing specifications. The first, c_i, specified the propensity of the examinees to guess while the second, c_g, specified whether guessing was to be accounted for in scoring.

Study 1: The effects of guessing

For this study the "perfect" item pool was used, with two values of c_g:$c_g = \begin{cases} 0 \\ .20 \end{cases}$, paired with two values of the personal guessing tendency $c_i = \begin{cases} 0 \\ .20 \end{cases}$. Of the four possible pairwise combinations, only three were used; resulting in three sets of conditions

	c_i	c_g
no guessing	0	0
uncorrected guessing	.20	0
corrected guessing	.20	.20

In the first condition, no guessing takes place (c_i=0) and no correction for guessing enters into the scoring formula (c_g=0). In the second condition c_i=.20 (every individual i has a random chance of correct response equal to .20), but c_g=0 (guessing goes uncorrected in the scoring algorithm). Finally, in the third condition, the .20 guessing parameter and the scoring correction for guessing take the same value.

In each condition, the same 100 "examinees" (θ_i sampled from a normal (0,1) population) were administered 14 simulated Bayesian sequential tests in which testing terminated for an examinee whenever the $\sigma^2{}_m$, the estimated variance of the posterior distribution of θ, fell below .0625 (this is equivalent to the Urry/Jensema criterion of SEE <.25). The 14 simulated tests in each condition were experimentally independent, and differed from each other in the value of the a_g parameter, which was constant within a test, but which varied systematically across tests. The 14 a_g values were a_g = .5, .6, .7, .8, .9, 1.0, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00.

For each test in each condition, the following variables were observed:

a. mean and range of test length, k
b. errors of estimate, $e_i = (\hat{\theta}_i - \theta_i)$
c. test bias, $(1/N) \sum_i (\hat{\theta}_i - \theta_i)$
d. mean absolute error, $(1/N) \sum_i |\hat{\theta}_i - \theta_i|$
e. test validity $r_{\theta\hat{\theta}}$
f. correlated error $r_{\hat{\theta}e}$ and $r_{\theta e}$
g. correlated test length $r_{\theta k}$ and $r_{\hat{\theta}k}$

Study 2: *The effects of the configuration of item parameters in the item pool*

Most simulation studies of Owen's sequential test have used a constant item discrimination parameter within each test. Typical item pools in actual use, however, have varying item discriminations, with the potential effect of having more discriminating items available in some ranges of the trait level than in others. In this study, different item pool a_gx b_g configurations were simulated by using the differentially discriminating "perfect" item pool. The approximate correlation (r_{ab}) between item discriminating power and item difficulty was varied in order to observe its effect on some properties of the Bayesian test and of the resulting scores.

Three different values of r_{ab} were simulated: -.71, 0 and +.71. With r_{ab}=.71, more discriminating items are available, on the average, at higher levels of θ.With r_{ab}=-.71 the more discriminating items were available at the lower levels of θ. And with r_{ab}=0, no level of θ was favored in terms of available discriminating power of the items, although discriminating power was free to vary randomly. In each "item pool" configuration, the mean item discrimination $\bar{a}_g$ was set at 1.25. Additionally, a minimum a_g value of .80 was imposed, in accord with Urry's (1974) recommendation.

The item pool configuration was simulated by means of:

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

2) calculating a conditional a_g value from a linear transform of $b^*{}_g$:

$$a_g|b^*{}_g = r_{ab} \left(\frac{\text{S.D.}_A}{\text{S.D.}_B}\right). b^*{}_g + \bar{a}_g$$

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_b, $b^*{}_g$, r_{ab}, $\bar{a}_g$ 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 population normal in $(0,\sigma^2{}_e)$

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

4) setting $a^*{}_g$ equal to .80 whenever it would otherwise have a lower value.

"Examinees" for this study were 3100 simulated θ's, 100 at each of 31 equally spaced intervals between -3.0

and 3.0, inclusive. The corrected guessing condition (c_g=c_i=.20) was in effect. The posterior variance termination criterion ($\sigma^2{}_m \leq .0625$) was used, with an arbitrary 30-item maximum test length. At each of the 31 θ levels the following variables were observed for each individual, i:

a. test length, k_i
b. test score, $\hat{\theta}_i$
c. error of estimate, $e_i = \hat{\theta}_i - \theta$

While study 1 examined average characteristics of the Bayesian test and test scores, Study 2 was concerned with certain properties of the procedure as a function of trait level, θ, and of the item pool configuration, r_{ab}. For each configuration, the regressions of k, e and $\hat{\theta}$ on θ were estimated from the means of the 100 individuals at each level of θ.

Additionally, the data were used to calculate empirical values of the information function $I_{\hat{\theta}}(\theta)$ of the Bayesian test scores θ. The information at any level θ_i may be calculated as the square of the ratio of the partial derivative with respect to θ of the regression of test scores $\hat{\theta}$ on θ, to the conditional standard deviation ($\sigma_{\hat{\theta}\,|\,\theta}$) of the test scores at the given level of θ. This may be written $I_{\hat{\theta}}(\theta) = \left[\frac{\partial/\partial\theta(\mathrm{E}(\hat{\theta}\,|\,\theta))}{\sigma_{\hat{\theta}\,|\,\theta}}\right]^2$ (after Lord, 1970, p. 153). In each configuration for each of the 31 levels of θ, the conditional standard deviation was estimated as the observed S.D. of the 100 test scores at that level. The numerator of the equation was calculated for each θ point from a third degree polynomial equation for the regression of $\hat{\theta}$ on θ, estimated by least squares fit to the thirty-one mean $\hat{\theta}$'s observed under each item pool configuration.

RESULTS

Study 1

Tables 1, 2 and 3 and Figures 1, 2 and 3 contain the results of sequential testing under the three conditions of guessing/correction for guessing, at each of 14 item discrimination levels. Some noteworthy trends are:

a. *Test length* was constant at each a_g level in the no guessing (Table 1; Figure 1) and uncorrected guessing (Table 2; Figure 2) conditions, with test length to termination diminishing proportionately with the inverse of the a_g level.

In the corrected guessing condition (Table 3 and Figure 3) test length varied across individuals, while *mean* test length within a_g level behaved in the same manner as did test length in the other two conditions. One datum of note is the behavior of test length as a function of a_g level: in order for all examinees to reach normal termination in less than 30 items (in the corrected guessing condition), the item discrimination value must exceed 1.25 (a_g>1.25).

Another result of interest is an expected one: the corrected guessing condition required more items to termination than did the other conditions.

b. *Errors of estimate,* $e_i = (\hat{\theta}_i - \theta_i)$, were moderately correlated with ability θ and test score $\hat{\theta}$ under all conditions, as revealed in Tables 1, 2 and 3. e_i tends to be positive for θ_i<0 and negative for θ_i>0. This result was consistent, and reflects a regression effect caused by the quadratic loss function employed in the item selection procedures.

c. Test bias, mean absolute error, test validity, correlated errors and correlated test length values for the no guessing, uncorrected guessing and corrected guessing conditions are listed in Table 1, 2 and 3, respectively. Additionally, Figures 1, 2 and 3 graph some of these values as a function of a_g level within each condition. Noteworthy in these data is the sizeable bias and mean absolute error in the uncorrected guessing condition (Table 2; Figure 2), as well as the tendency for bias and absolute error to increase at a_g levels above 2.00 in the corrected guessing condition (Table 3; Figure 3). Note also that in the uncorrected guessing condition (Table 2), test validity, $r_{\hat{\theta}\theta}$, decreased at a_g levels beyond 2.00. Jensema (1972) observed this phenomenon, which he termed "correlation drop-off."

Study 2

Table 4 lists the observed mean values under each item pool configuration of test score, test length, and error of estimate for each value of θ. Figures 4, 5 and 6 depict these data graphically.

a. *Test length.* Mean test length (Figure 4) did not vary with θ in the r_{ab}0 configuration since the maximum of 30 items occurred at all levels. In the r_{ab}-.71 configuration, mean test length covaried positively and almost perfectly with ability level. In the r_{ab}+.71 configuration, test length covaried inversely with trait level, with more items required at the lower trait levels until the arbitrary 30-item limit was reached.

b. *Test scores.* The regression of mean trait estimates, $\hat{\theta}$ on θ was virtually linear in all three configurations in the interval $[-1.5<\theta<2.0]$. As can be seen from Figure 5, the Bayesian test scores tended to underestimate θ at high trait levels, and to overestimate θ at low trait levels. The regression of $\hat{\theta}$ on θ departed from a linear regression at extreme levels of θ (beyond $\theta = \pm 2.00$) with the departure more noticeable in the lower extremes of the scale.

c. *Errors of estimate.* The regression of mean errors of estimate on θ, seen in Figure 6, clearly illustrates a tendency of the Bayesian test scores to overestimate θ markedly and consistently at θ<-1.5 in all three item pool configurations. The tendency to underestimate high θ's is also illustrated. In this data the latter tendency was quite strong with r_{ab}-.71 but less so with r_{ab}+.71.

Information. The estimated values of the derivative $\frac{\partial}{\partial\theta}[\mathrm{E}(\hat{\theta}\,|\theta)]$, the conditional standard deviation $\sigma_{\hat{\theta}\,|\theta}$ and the information at each level of θ, under each item pool configuration, are listed in Table 5. Smoothed information curves for all three configurations are plotted in Figure 7. Some noteworthy trends are pointed out here.

TABLE 1

Test Length, Mean Errors of Estimate, and Correlates of Ability θ and Test Score θ, as a Function of Item Discrimination a_g in the Perfect Item Pool. No Guessing Condition (c_g=c_i=0).

Property	Item Discrimination (a_g) .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
Error 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
$r_{\theta e}$	-.17	-.08	-.10	-.16	-.20	-.15	-.17	-.14	-.07	-.15	-.16	-.14	-.09	-.10
with test length														
$r_{\theta k}$	...[a]	...	...	...	...	...	...	...	...	...	...	...	...	...
$r_{\theta k}$	...	...	...	...	...	...	...	...	...	...	...	...	...	...
$r_{\theta\hat{\theta}}$(validity)	.98	.98	.98	.98	.98	.98	.98	.97	.97	.97	.97	.97	.97	.97

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

TABLE 2

Observed Properties of the Bayesian Sequential Test as a Function of Item Discrimination in the Perfect Item Pool. Uncorrected Guessing (c_g=0; c_i=.20)

Property	Item Discrimination (a_g) .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	.05	.05	.05	.07	.02
with test length	[a]													
$r_{\theta k}$	....	....	....	....	....	....	....	....	....	....	....	....	....	....
$r_{\hat{\theta} k}$	....	....	....	....	....	....	....	....	....	....	....	....	....	....
$r_{\theta\hat{\theta}}$(validity)	.97	.97	.96	.95	.95	.95	.96	.94	.95	.93	.93	.93	.92	.91

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

TABLE 3

Observed Properties of the Bayesian Sequential Test as a Function of Item Discrimination in the Perfect Item Pool. Corrected Guessing (c_g=c_i=.20)

Property	Item Discrimination (a_g)													
	.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														
$r_{\theta e}$	-.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
$r_{\theta k}$	....[a]	.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
$r_{\theta\hat{\theta}}$	.97	.98	.99	.98	.98	.98	.98	.98	.98	.98	.98	.97	.97	.97

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

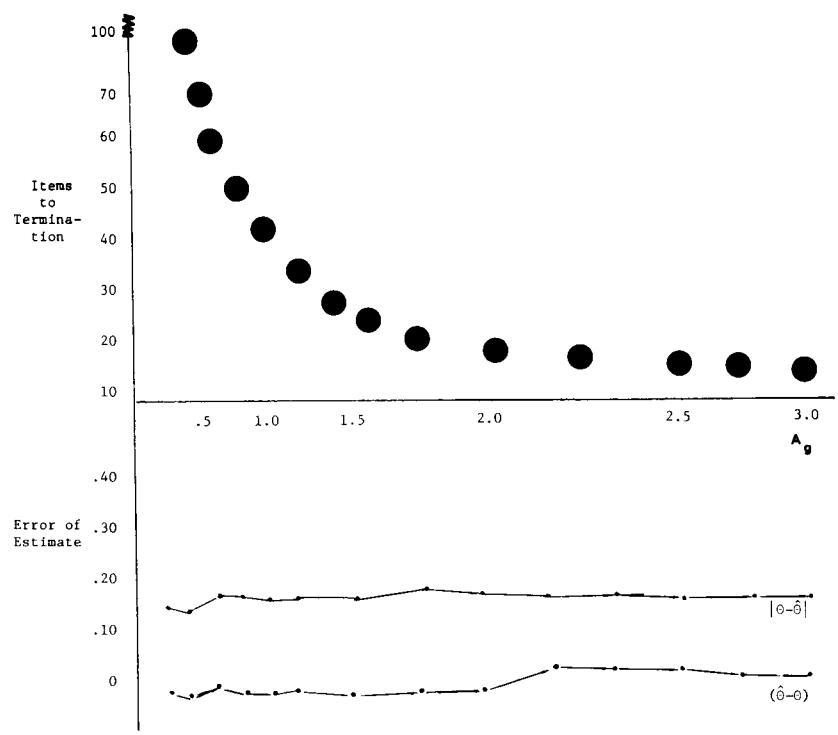

Figure 1. Some observed properties of a Bayesian sequential test, as a function of item discrimination. No guessing; perfect item pool; posterior variance termination criterion.

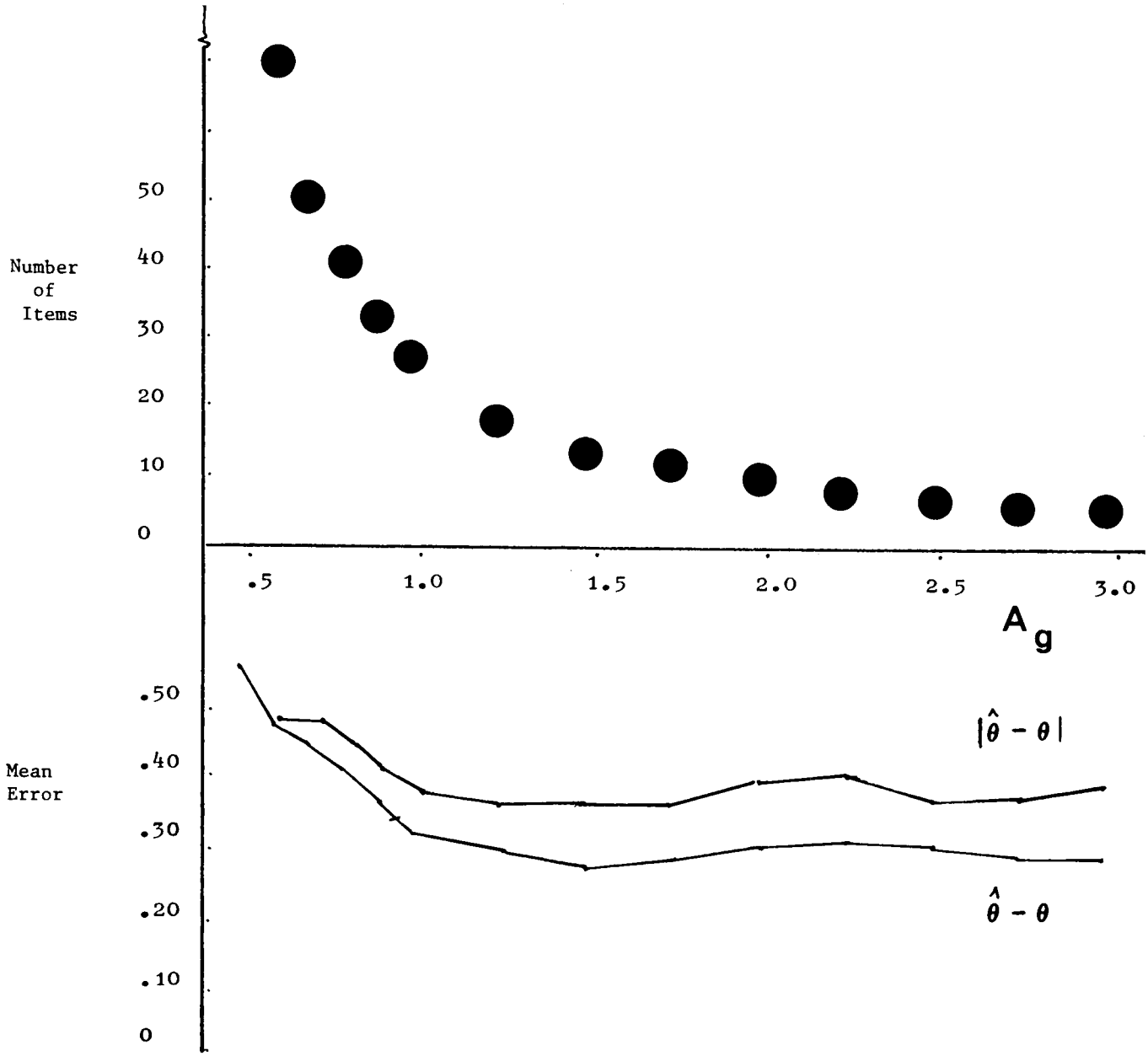

Figure 2. Some observed properties of a Bayesian sequential test, as a function of item discrimination. Uncorrected .20 guessing; perfect item pool; posterior variance termination criterion.

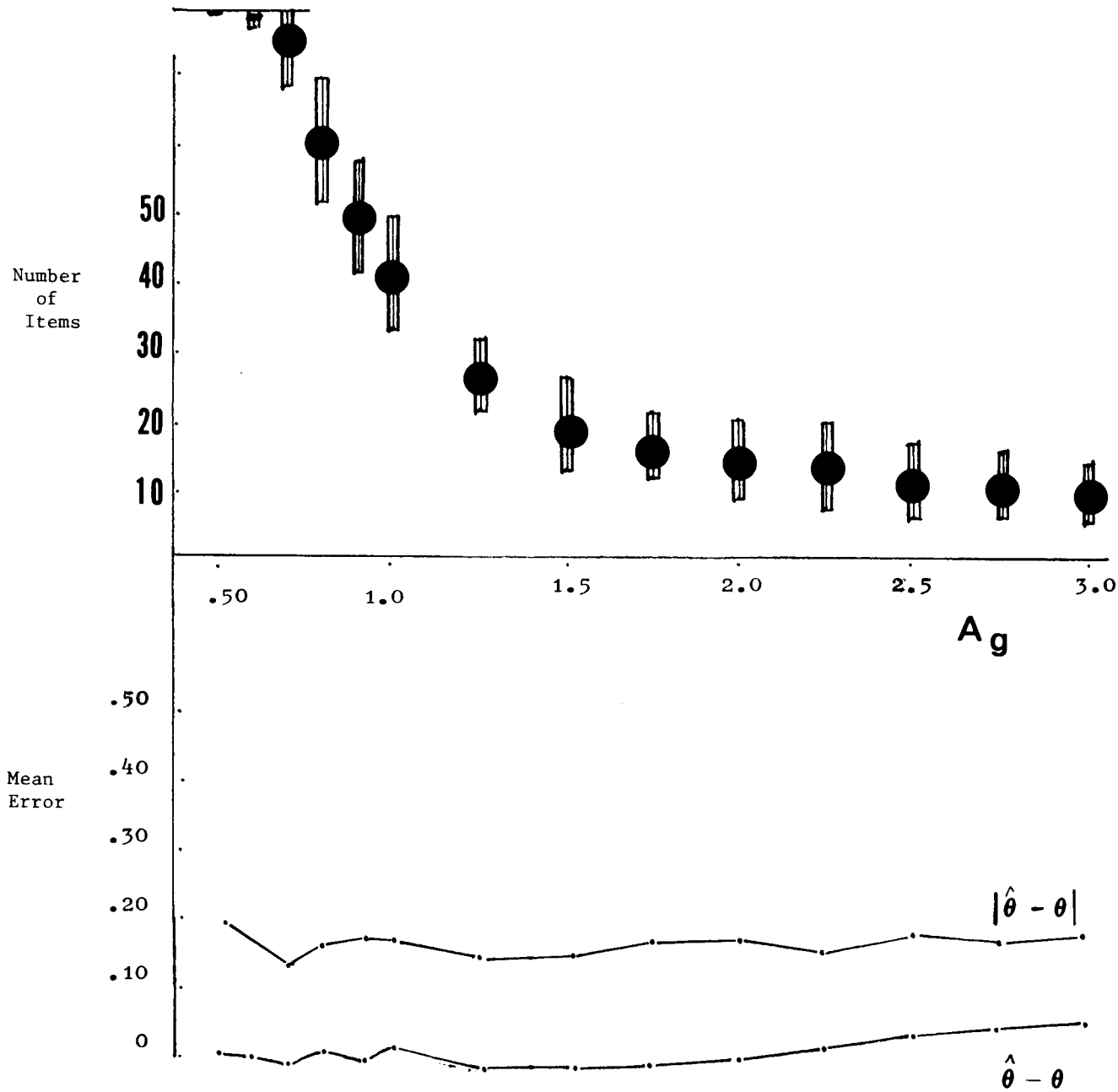

Figure 3. Some observed properties of a Bayesian sequential test, as a function of item discrimination. Corrected .20 guessing; perfect item pool; posterior variance termination criterion.

TABLE 4

Mean Test Scores ($\hat{\theta}$), Mean Test Length (k) and Mean Error of Estimate (e) for Three Item Pool Configurations, at each of 31 Trait Levels (θ)

	Item Pool Configurations								
	r_{ab}+.71			r_{ab}.0			r_{ab}-.71		
θ	$\hat{\theta}$	k	e	$\hat{\theta}$	k	e	$\hat{\theta}$	k	e
-3.0	-2.39	30	.612	-2.47	30	.532	-2.30	14	.696
-2.8	-2.26	30	.545	-2.29	30	.513	-2.20	14	.601
-2.6	-2.06	30	.542	-2.25	30	.352	-2.17	15	.427
-2.4	-2.00	30	.404	-2.06	30	.342	-2.08	15	.317
-2.2	-1.81	30	.390	-1.94	30	.263	-1.93	16	.269
-2.0	-1.70	30	.296	-1.80	30	.204	-1.74	17	.263
-1.8	-1.60	30	.200	-1.66	30	.141	-1.65	18	.146
-1.6	-1.44	30	.163	-1.45	30	.151	-1.48	18	.125
-1.4	-1.24	30	.162	-1.32	30	.082	-1.29	20	.110
-1.2	-1.12	30	.076	-1.12	30	.082	-1.14	21	.060
-1.0	- .93	30	.073	- .93	30	.071	- .98	22	.018
- .8	- .74	30	.055	- .74	30	.055	- .76	24	.037
- .6	- .56	30	.038	- .59	30	.014	- .58	26	.015
- .4	- .44	30	-.040	- .40	30	.004	- .35	27	.049
- .2	- .25	30	-.046	- .21	30	-.010	- .14	29	.062
0	- .06	30	-.058	.05	30	.046	.02	30	.021
.2	.20	30	-.003	.16	30	-.039	.19	30	-.007
.4	.35	30	-.053	.34	30	-.056	.35	30	-.051
.6	.53	29	-.068	.61	30	.010	.58	30	-.015
.8	.76	29	-.044	.74	30	-.058	.81	30	.013
1.0	.95	28	-.051	.89	30	-.113	.92	30	-.080
1.2	1.11	27	-.091	1.16	30	-.036	1.15	30	-.047
1.4	1.37	26	-.034	1.33	30	-.068	1.25	30	-.150
1.6	1.53	26	-.074	1.48	30	-.117	1.46	30	-.140
1.8	1.73	25	-.070	1.68	30	-.123	1.64	30	-.165
2.0	1.89	24	-.113	1.88	30	-.119	1.78	30	-.224
2.2	2.09	24	-.107	2.05	30	-.146	1.98	30	-.224
2.4	2.27	23	-.132	2.22	30	-.176	2.13	30	-.270
2.6	2.47	23	-.126	2.37	30	-.230	2.33	30	-.273
2.8	2.63	23	-.168	2.57	30	-.230	2.43	30	-.372
3.0	2.81	23	-.189	2.72	30	-.282	2.57	30	-.426

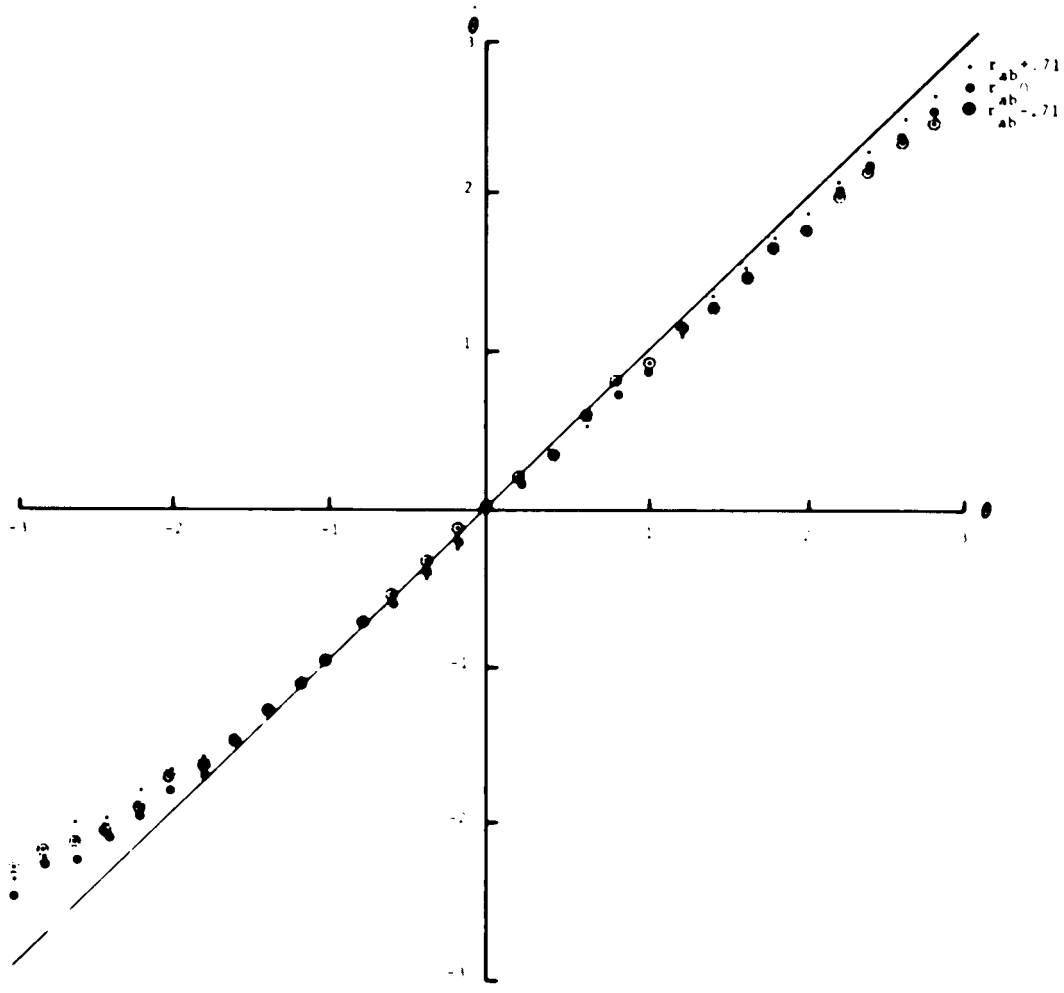

Figure 4. Mean estimated ability ($\hat{\theta}$) at thirty-one ability points (θ) for the simulated Bayesian sequential test under three item pool configurations.

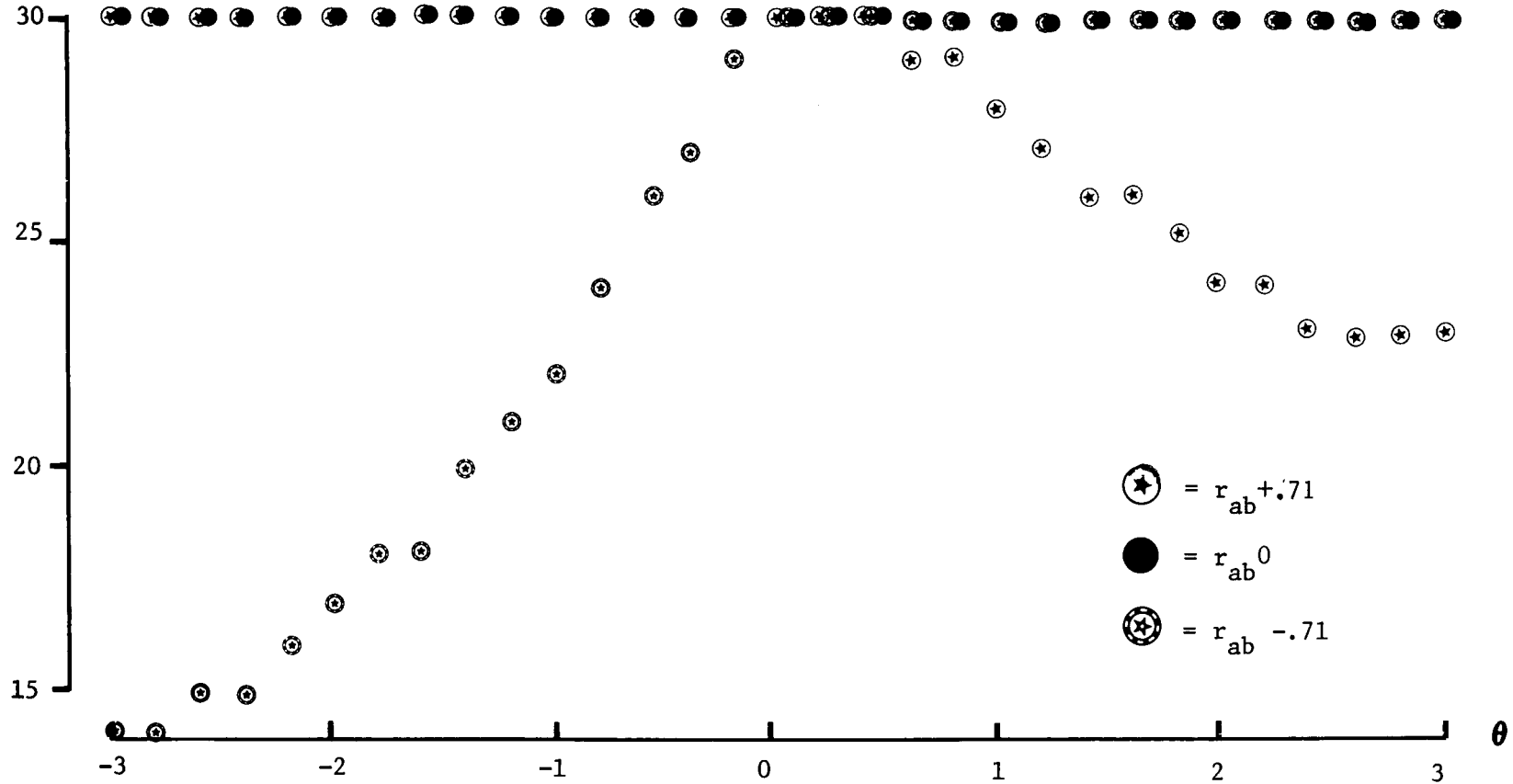

Figure 5. Mean number of items to termination (test length) at thirty-one ability points (θ) for the simulated sequential test under three item pool configurations (See text.)

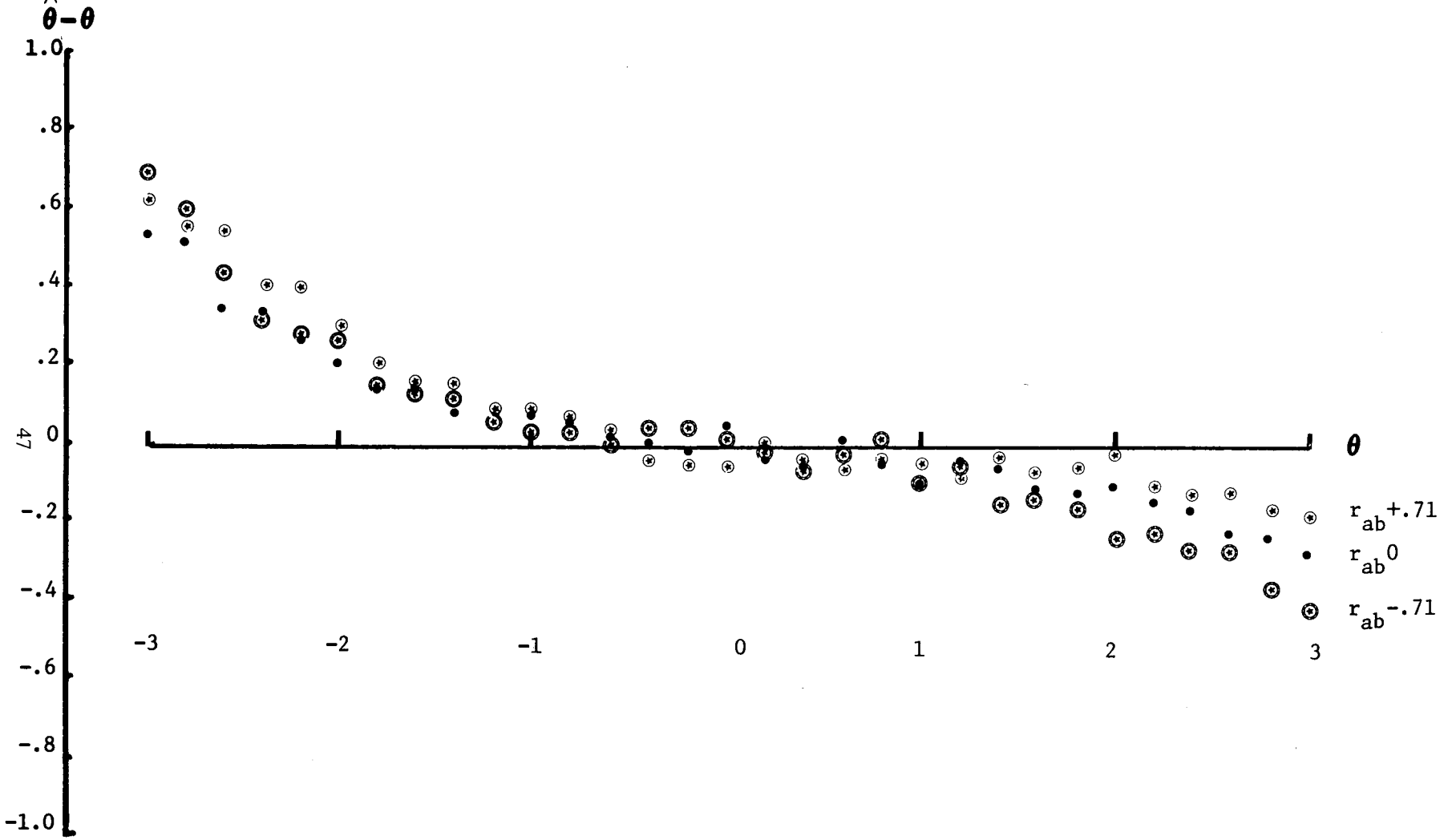

Figure 6. Mean error of estimate $\hat{\theta}-\theta$) at thirty-one ability points (θ) for the simulated Bayesian sequential test under three item pool configurations.

TABLE 5

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 31 Ability 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	.942	.255	14.66
- .2	.927	.308	9.06	.946	.278	11.59	.953	.284	13.96
0	.940	.305	9.50	.954	.249	14.68	.960	.257	13.96
.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

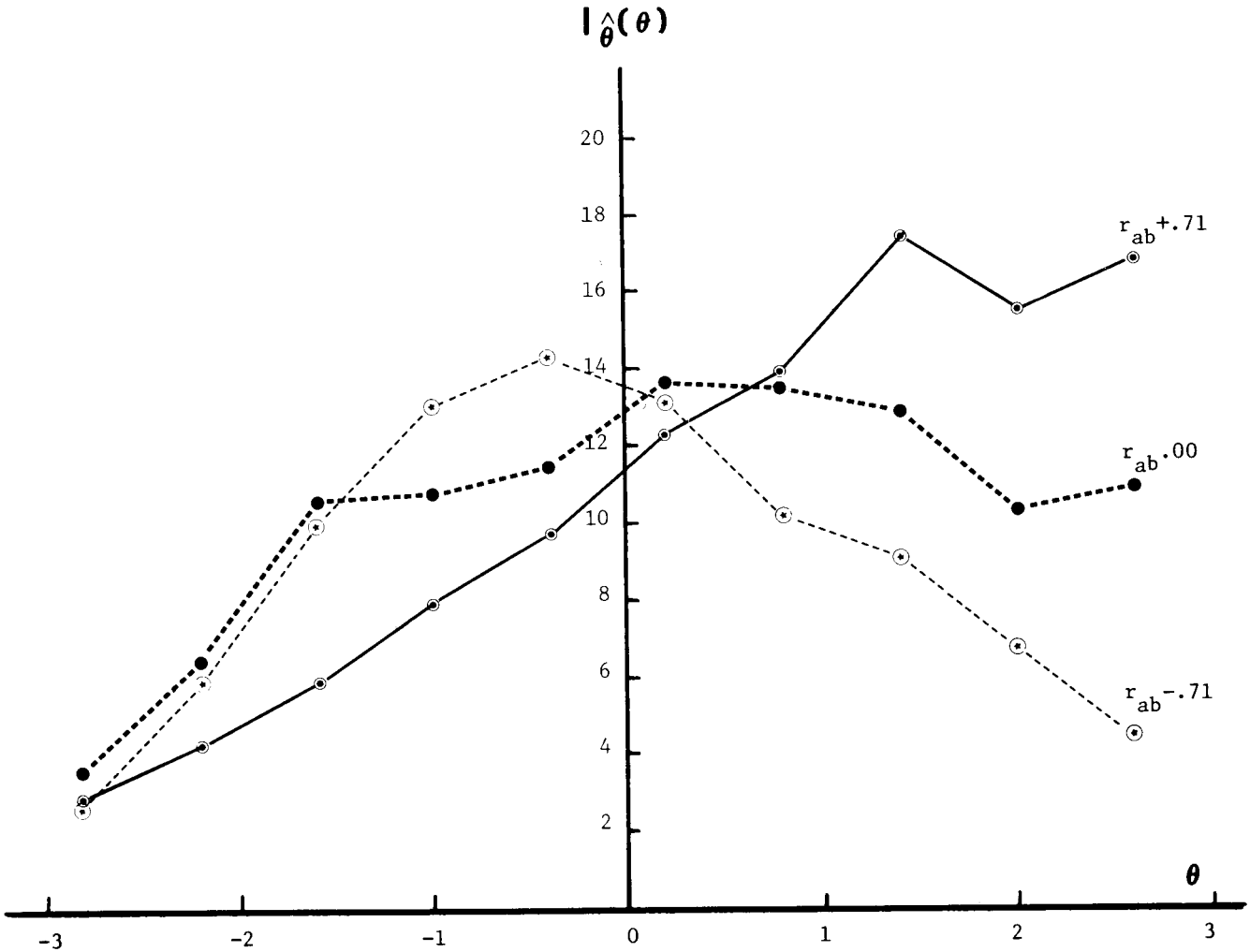

Figure 7. Smoothed curves of the information functions of the Bayesian sequential test under three different item pool difficulty-by-discrimination configurations. (see text.)

1) Under all three item pool configurations the information functions were very low in the low end of the θ distribution;

2) For r_{ab}+.71 the information values uniformly increased with increasing θ;

3) For r_{ab}0 information generally increased with θ, to about θ = 1.00, then decreased somewhat;

4) For r_{ab}- .71 information increased sharply with θ, to about θ = 0, then just as sharply decreased.

DISCUSSION

Study 1

Test length, or number of items required to satisfy the posterior variance termination criterion, was shown to vary inversely with item discriminatory power, a_g, when the latter is constant for all items in a given test. This result was expected, and corroborates the findings of Jensema (1972, 1974) who also pointed out that if constant item discriminatory powers were available it would be possible to predict the validity of the trait estimates from the number of items administered, and conversely to estimate the number of items required to achieve any given validity level.

In the no-guessing and uncorrected guessing conditions (that is, in tests which assume no guessing) the test length was constant for any fixed a_g value. This result would not be likely to occur with a finite pool of items due to the inevitability of imperfect θ-with-item-difficulty matches. That is, with a finite item pool some variance in test length would likely occur even if all items had equal discrimination parameters. The fact that there was no variance in test length (within any given discrimination level) with the perfect item pool indicates that any variance in test length in a real, constant-discrimination, no-guessing test must be due solely to inadequacies in the distribution of item difficulty parameters in the finite item pool.

These results are pertinent to the use of Rasch-model ability estimation in an adaptive testing situation. Except for the specification of the item characteristic function, the Rasch model is conceptually identical with the no-guessing model used in Study 1. Within each test, item discrimination parameters were constant (as the Rasch model assumes) and no-guessing was assumed. Thus the major difference between this portion of Study 1 and a Rasch model simulation would be in the definition of the item response model. We assumed a one-parameter normal ogive response model, whereas the Rasch model uses a one-parameter logistic one (Birnbaum, 1968, p. 402). As Birnbaum (1968, p. 399) has pointed out, the two response models are very similar. Thus, the results of Study 1 for the no-guessing condition should be generalizable to adaptive tests based on the Rasch model.

In the corrected guessing condition (Figure 3) there was some variance in test length for all a_g values (except a_g = .50, where no testees terminated in fewer than 100 items). For all a_g levels above .50, test length θ correlated strongly and positively with the trait estimate $\hat{\theta}$ (Table 3). The test length -$\hat{\theta}$ correlation $r_{\hat{\theta}k}$ equalled or exceeded .80 for all a_g values above .6. The correlation $r_{\theta k}$ between test length and ability θ was of similar magnitude but always smaller than $r_{\hat{\theta}k}$. It seems obvious that for the case of constant item discrimination and non-zero guessing there is a systematic relationship between ability θ or test score $\hat{\theta}$ and number of items administered. Examination of the partial correlations, however, shows that $r_{\theta k}$ vanishes when θ is statistically controlled for. For instance, for a_g = 1.0 we observed $r_{\theta k}$ = .81, $r_{\hat{\theta}k}$ = .83, $r_{\theta\hat{\theta}}$ = .98. Controlling for $\hat{\theta}$ and θ, respectively, yields the following partial correlations:

$$r_{\theta k.\hat{\theta}} = -.03$$

$$r_{\hat{\theta}k.\theta} = .31$$

Analysis of the corresponding partial correlations for the other a_g levels would yield a similar result: $r_{\theta k.\hat{\theta}}$ approximately zero, but $r_{\hat{\theta}k.\theta}$ positive and moderate. This suggests that, at least for the constant item discrimination case, the tendency for $r_{\hat{\theta}k}$ to be positive is due to some characteristic of the trait estimation method using the guessing correction.

Another observation with regard to test length has a practical application. Where the posterior variance termination criterion is to be used, it is desirable that all or nearly all examinees reach criterion (e.g., $\sigma^2_m \leq .0625$ or some other arbitrary value) within a reasonably small number of items. Typically (e.g., Urry, Jensema), a 30-item maximum test length has been imposed in conjunction with the posterior variance criterion. If a large number of examinees reach the 30-item limit before attaining the posterior variance criterion, the latter may lose its usefulness as a predictor of test validity. The data of Table 3 (and Figure 3) indicate that even with a "perfect" item pool, the constant item discrimination parameter must equal or exceed a_g = 1.25 in order to insure test termination in fewer than 30 items for the majority of examinees when guessing is a factor. Although it is difficult to generalize this finding to the case of typical finite item pools, it is reasonable to expect that test termination via the posterior variance criterion $\sigma^2_m < .0625$ will seldom occur in fewer than 30 items in Bayesian sequential tests using item pools whose mean item discrimination parameter is less than 1.25.

Errors of estimate were moderately and negatively correlated with θ in all three conditions, with the strongest correlations observed in the uncorrected guessing situation. That is, with constant item discrimination and a perfect pool of item difficulties, larger errors of estimate ($\hat{\theta}$-θ) tended to occur as θ decreased. This tendency can be viewed as a regression effect. As is typical with linear regression estimates for all three conditions the estimates $\hat{\theta}$ tended to be closer to the mean than the actual values θ.

The correlation $r_{\hat{\theta}e}$ between trait estimates $\hat{\theta}$ and errors $(\hat{\theta}-\theta)$ was consistently of the same sign but lower magnitude than $r_{\theta e}$, with the no guessing and corrected guessing conditions.

The mean error of estimate, or bias, was virtually zero in the no guessing condition, until a_g became large (Table 1; Figure 1). For $a_g \geqslant 1.50$ there was a tendency for positive bias to occur. Similarly, mean absolute error was quite constant until a_g=1.50, than became larger. In the corrected guessing condition (Table 3, Figure 3) mean absolute error was fairly constant across a_g levels, but bias was positive at low a_g values, diminished virtually to zero at intermediate levels, and began to increase steadily as a_g increased above 2.0.

Study 2

Test length. The data illustrate clearly the effect of item pool configuration on the correlation of test length with θ (or $\hat{\theta}$): The correlation is strong and its sign was opposite that of the r_{ab} correlation in the simulated item pool. (For the $r_{ab}0$ configuration there was no variance in test length, due to the arbitrary 30-item limit. The preceding three studies have shown, however, that with constant a_g, test length varies directly with θ. Presumably that relationship would hold for the $r_{ab}0$ configuration if test length was free to exceed 30 items). We have already alluded to the inverse relationship between test length and the rate of reduction in the Bayes posterior variance. Thus, it should be clear that the configuration of difficulty and discrimination parameters in the item pool, which can be roughly described by the correlation of the discrimination and difficulty parameters (r_{ab}), effectively dictates the rate of posterior variance reduction at any level of the trait θ. Furthermore, if a maximum test length is arbitrarily established (such as the 30-item limit used by us, and by Urry, 1974, and Jensema, 1972) that limit, in conjunction with the item pool configuration, may dictate regions of the θ continuum in which satisfactory convergence of the trait estimates will seldom occur.

Errors of estimate. Study 1 found very high validities of the trait estimates $\hat{\theta}$, indicating that the Bayesian sequential test is capable of ordering simulated examinees from a normal population quite well with respect to the variable, θ, underlying the item responses. Study 2 was motivated by an interest in the *accuracy* of the estimates of θ, rather than the correctness of ordering, as a function of θ itself. The data showed clearly that the Bayesian estimates behaved in a manner similar to linear regression, except at the extremes of the normal distribution ($\theta \leqslant -1.5$ and $\theta \geqslant 2.0$). Typically, linear regression underestimates the criterion variable above the mean, and overestimates it for values below the mean. Such was the case for the Bayesian sequential estimates, except that the underestimates became fairly sizeable (around .20) on the average for $\theta > 2.0$, and overestimates became severe (larger than .5) in the lower levels of the trait. Furthermore, it was shown that the behavior of the trait estimates varies as a function of the item pool configuration. Thus, by controlling the item pool configuration for a live-testing item pool it should be possible to control the accuracy of the Bayesian test scores as estimators of the actual trait level of the examinees. Other alternatives may prove useful in this regard. Some of these will be discussed below.

Information. For the configuration r_{ab}+.71, the information of the trait estimates appears to increase linearly with θ, at least in the interval $[-3.0 \leqslant \theta \leqslant 3.0]$. This is what we might expect, since item discrimination increased with θ in this configuration. Note (Table 4) that mean test length in this configuration was 30 items for $-3 \leqslant \theta \leqslant .6$, and then decreased linearly with for $\theta < .6$, reaching a mean of 23 items at $\theta = 3.0$.

For the $r_{ab}0$ configuration the information function appeared to take the shape of an inverted (and rather asymmetric) shallow dish, with maximal information attained in the interval $[0 \leqslant \theta \leqslant 1.5]$. This should approximate, at least in its form, the information structure resulting from applying the Bayesian sequential test with a real item pool whose configuration is based on Urry's (1974) prescription. It should be apparent that some efficiency of measurement will be lost in the extremes of the θ distribution, especially in the lower extremes. Note that for these data, test length was a constant 30 items at all levels.

For the r_{ab}-.71 configuration the information curve does not take the shape one would assume intuitively. From knowledge of the distribution of the discrimination parameters it would seem that the curve should mirror that of the r_{ab}+.71 information but with maximal information at θ=-3.0. Instead it rather emphatically takes the convex form. The test is maximally efficient in the interval $[-1 \leqslant \theta \leqslant 0]$, and rapidly loses efficiency elsewhere. This is a remarkably different result from what one would expect. The highest item discrimination parameters were available at the low end of the θ scale, yet information was as low there $[-2 < \theta < -1.5]$ as it was where the lowest item discrimination values occurred $[1.5 < \theta < 3.0]$. The low levels of information in the low θ region are due in part to the small number of items administered there. As Table 4 reveals, the posterior variance termination criterion resulted in mean test length of 14 items at θ=-3.0; 17 items at θ=-2.0; 22 items at θ=-1.0 The information values obtained with these test lengths could be adjusted statistically to estimate the information values for constant 30 item test length. Such an adjustment would still show an efficiency loss at $\theta < -2.0$ for this item pool configuration, despite the high average item discrimination in that region. We will address this problem further in the discussion to follow.

Implications. These results were obtained by simulating a "perfect" item pool; i.e., a pool in which unlimited numbers of items of any difficulty level were available. This should result in data, which, within the limits of sampling error, approximate the best possible results obtainable using the sequential testing procedure as specified by Owen (1969), under the conditions studied.

We have found, as did Urry (1971, 1974) and Jensema (1972, 1974) before us, that the procedure has the potential to yield trait estimates having very high validities with great economy in test length, provided that highly discriminating test items, rectangularly distributed on difficulty, consitute the item pool. We have also found that there may be a tendency of the method to overestimate group mean trait level, when item discrimination parameters are very high, even when the trait estimation model exactly conforms to the item response model. When the estimation model is not congruent with the item response model (as in the uncorrected guessing condition of study 1) we have found that rather sizable bias of estimate may occur, accompanied by diminished validity.

Lord (1970, p. 152) made the point that evaluating a tailored test by means of a group statistic (such as our validity coefficient $r_{\theta\hat{\theta}}$) 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 was, as we have seen, quite high under the conditions used in our 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 functions of the trait estimates showed that the effectiveness of measurement under the Bayesian tailoring 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 the advantage of tailored testing over conventional testing is the former's supposed potential for superior measurement accuracy and effectiveness in those extremes. From our data it is apparent that with the exception of the r_{ab}+.71 configuration, the sequential test scores are behaving much like conventional test scores, at least in terms of the shapes of their information functions. And even for the r_{ab}-.71 configuration measurement effectiveness was relatively poor in the lower extremes of θ. The utility of the Bayesian adaptive testing strategy may be diminished considerably by results like those reported for Study 2, if they prove to be general.

The problems revealed in Study 2 (of bias non-linear in θ, and of convex information structures of the trait estimates) have causes which may be amenable to improvement. At the heart of the problem 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 problem is the Bayesian procedure itself. As we have 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. Now, the more extreme θ is for any individual, the larger will be the regression effect, on the average. Recall that the item selection procedure selects an item with difficulty b_g 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, say 30 items, the effects of this inappropriate item selection will be several:

1) mean proportion correct will tend to increase as a function of θ, despite the explicit 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 $\hat{\theta}$ or θ.

For low θ the initial prior is an overestimate. Hence, the first item selected will generally be too difficult $[(b_g-\theta)>0]$, yet the examinee has a non-zero chance of answering it correctly. A correct answer, of course, will cause an increase of $\hat{\theta}$ and thus result in another inappropriate choice of item difficulty. Furthermore, as Samejima (1973) has shown, there may actually be negative information in a correct response to an item whose difficulty b_g exceeds an examinee's actual trait level θ by a fairly small increment, when guessing is a factor. We suggest that examinees in the low extremes of θ are rather consistently being administered overly difficult items $[(b_g-\theta)>0]$ 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, we have not looked at proportion correct as a function of θ, nor at the quantity $(b_g-\theta)$, both of which bear on the appropriateness of the tailoring process. Future simulation studies will be necessary to examine these variables.

One goal of adaptive testing should be to achieve a constant high level of measurement effectiveness at all levels of θ. This desideratum is equivalent to a high, horizontal information function. We have found that the Bayesian sequential test failed to achieve this goal despite an unrealistically favorable set of circumstances: the perfect item pool, errorfree item parameters, and a scoring model perfectly congruent with the item response model. We have attributed the shortcomings of the Bayesian trait estimates to the regression-like tendency of the sequential

estimates themselves, which in turn result in inappropriate item selection for individuals whose trait levels are extremely high or low.

There are at least two methods of ameliorating this problem, both of which should, to some extent, lessen the bias of estimate at the extremes and improve the information structure 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 present item selection procedure with a mechanical 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 do considerable violence to Owen's elegant procedure.

If the practitioner is committed to the procedure as it was originally proposed, it would seem that the best course of action would be to take great care in assembling the item pool, and to administer a constant number of items (say 30) to each examinee. If no strong commitment to Owen's procedure is involved, the practitioner may be well advised to use another adaptive strategy, such as Weiss' stradaptive test (Weiss, 1974), Lord's (1974) maximum likelihood procedure, or a similar procedure being investigated by Samejima (1975). Systematic investigation of some of these strategies, which will permit them to be compared with the Bayesian sequential test, are currently in progress.

REFERENCES

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).

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. 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. & Weiss, D. J. Simulation studies of Bayesian adaptive ability testing, 1975 a. (In preparation.)

McBride, J. R. & Weiss, D. J. An empirical study of Bayesian computerized testing, 1975b. (In preparation.)

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, in press.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika* 1973,*38*(2), 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.

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

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