Bias and Information of Bayesian Adaptive Testing

David J. Weiss
University of Minnesota

James R. McBride
Navy Personnel Research and Development Center

Monte carlo simulation was used to investigate score bias and information characteristics of Owen's Bayesian adaptive testing strategy and to examine possible causes of score bias. Factors investigated in three related studies included effects of an accurate prior θ estimate, effects of item discrimination, and effects of fixed versus variable test length. Data were generated from a three-parameter logistic model for 3,100 simulees in each of eight data sets, and Bayesian adaptive tests were administered, drawing items from a "perfect" item pool. Results showed that the Bayesian adaptive test yielded unbiased θ estimates and relatively flat information functions only in the situation in which an accurate prior θ estimate was used. When a constant prior θ estimate was used with a fixed test length, severe bias was observed that varied with item discrimination. A different pattern of bias was observed with variable test length and a constant prior. Information curves for the constant prior conditions generally became more peaked and asymmetric with increasing item discrimination. In the variable test length condition, the test length required to achieve a specified level of the posterior variance of θ estimates was an increasing function of θ level. These results indicate that θ estimates from Owen's Bayesian adaptive testing method are affected by the prior θ estimate used and that the method does not provide measurements that are unbiased and equiprecise except when an accurate prior θ estimate is used.

Since test scores are typically used to differentiate among persons, one highly desirable property of a test is that it measure equally well at all levels of a trait. Another consideration is that it measure each person precisely. Thus, an "ideal" test would have a high horizontal information function. Unfortunately, this ideal cannot normally be achieved in a fixed-length conventional test that is constructed from a much larger fixed pool of test items. Ordinarily, some tradeoffs must be made. Relatively high information at a point can be achieved by "peaking" the test, that is, constructing it of the most discriminating items in a narrow range of difficulty. A relatively flat but low information function can be achieved by selecting equidiscriminating items having a wide range of item difficulty values. The only way to approximate a high flat information function is to administer to each person the subset of items that provides the most information at his/her trait level, θ. The problem with this is obvious: θ is unknown before the test is administered.

An adaptive test can select items during the course of testing in such a way as to attempt to maximize the information obtained for each examinee. This may be done either by simple branching—administering a more difficult item after a correct answer and an easier item after an incorrect answer—or by more elaborate techniques (Weiss, 1982). Owen's (1969, 1975) Bayesian adaptive testing strategy estimates θ after each item response, then selects the unused test item that is, in one sense, the most "informative" at the current estimated θ level.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 3, Summer 1984, pp. 273–285

0146-6216/84/030273-13$1.90

The result is that different persons take different sets of test items; each set of test items spans a range of difficulty levels approximately tailored to provide maximal information about the individual examinee.

The information function of the test scores derived from any adaptive testing procedure should be (1) flatter than that of a peaked test of the same length and constructed from the same item pool and (2) higher than that of a rectangular test of the same length drawn from the same item pool. The height of the adaptive test's information function will be determined in large part by the discrimination and guessing parameters of the constituent items of the item pool as well as by test length. The flatness of the information curve (and to some extent its height) will depend largely on the range of item difficulties in the pool and on the effectiveness of the adaptive item selection procedure.

Urry (1971) conducted monte carlo simulations of Owen's (1969, 1975) adaptive testing procedure using three different simulated item banks: two banks of "ideal" item parameters and one bank of items with the same parameters as the Verbal Scholastic Aptitude Test (VSAT; Lord, 1968). Urry's Item Bank A had 20 equidiscriminating items ($a = 1.6$) at each of five equally spaced levels on the θ continuum. His Item Bank B employed five items of the same ($a = 1.6$) discriminations at each of 20 θ levels, and Item Bank C employed the parameters actually occurring in the VSAT. Banks A and B required an average of just over 11 items to test termination. Bank C required an average of 27.5 items to termination. The other noteworthy result of Urry's (1971) simulation studies was the magnitude of the fidelity coefficients. For simulated examinees drawn randomly from a normal (0,1) population, the observed correlations of .936 (Item Bank A) and .919 (Item Bank B) are quite high in view of the relatively short test lengths involved.

Jensema (1972) simulated Owen's (1969, 1975) approach to Bayesian adaptive testing using the actual item responses of 100 live examinees to 58 mathematics items drawn from four conventional precollege tests taken at full length by the examinees. From a record of their item-by-item actual test performance, a computer program constructed artificial protocols of their responses to the items that would have been administered by a Bayesian adaptive test under two different conditions: with and without differential prior information about examinees' abilities. Parallel to these two "real data" simulations, Jensema carried out monte carlo simulations of the Bayesian procedure. These simulations used 100 simulated examinees and items with logistic ogive parameters identical to the 58 real items. Item scores were generated as a stochastic function of θ and of the parameters of each item. The adaptive tests were terminated in each instance when the posterior variance of the Bayesian θ estimate fell below .0625 or when 30 items had been administered, whichever occurred first.

In the real-data simulation, mean test length was about 27 items, with or without differential initial θ estimates. The Bayesian estimates correlated about .86 with scores on a weighted composite of the four conventional tests from which the item bank was selected. Jensema did not report a correlation of ability with test length or with precision of estimate, but he did observe that the posterior variance criterion terminated the testing only in the upper portions of the distribution of estimated ability. Jensema interpreted these results to imply that the item pool was unsatisfactory for adaptive testing in the lower ability levels due to the low discriminations of the items in that region of the difficulty continuum. His monte carlo results using the same item pool resulted in virtually identical mean test lengths and in correlations of .92 between estimated and true θ. He concluded, in part, that a satisfactory item pool for adaptive testing needs to employ very highly discriminating items uniformly distributed on the difficulty continuum. Another conclusion he reached—this one on the basis of monte carlo simulation with ideal item banks—was that for most purposes little was to be gained by the use of prior information about examinees to determine a variable initial θ estimate. Jensema found that using differential prior information resulted in an average savings of only one test item.

In another monte carlo study of Owen's Bayesian strategy, Jensema (1974) examined the effects of item parameters and test length on test reliabil-

ity. He showed that reliability is directly related to the posterior variance of the Bayesian θ estimate; hence, using a specific value of the posterior variance as a termination criterion determines the reliability of the test. Jensema showed that the average number of items required to attain a specified reliability varies as a function of the item parameters. With items uniformly distributed in difficulty, the higher the item discrimination, the shorter the test.

McBride (1977; McBride & Weiss, 1976) also studied characteristics of the θ estimates resulting from Owen's (1969, 1975) strategy. These monte carlo simulations involved (1) an ideal item pool with variable test length; (2) the effects of guessing and item discrimination in a perfect item pool; (3) the effects of fixed test length; and (4) the effects of θ level and item pool configuration. In the first three studies, the performance of the adaptive test was evaluated on overall indices that included the overall bias and mean absolute error of the θ estimates, the correlation of θ estimates with true θ (fidelity), and the correlations of true and estimated θ levels with errors and test length.

The fourth study evaluated the performance of this testing strategy in an item pool with no correlation between difficulty and discrimination parameters, and using items with high negative and high positive correlations between these parameters. In contrast to the other studies, characteristics of the θ estimates were examined as a function of true θ; dependent variables included bias and information conditional on θ. Contrasting with the first three studies, which showed little overall mean bias and information, Study 4 showed severe bias in the conditional θ estimates for all three item pool configurations. Estimates of θ were unbiased only for five θ values between $\theta = 1.0$ to -1.0; for low θ values θ was overestimated, and high θ values were underestimated. In addition, the information curves for the three item pool configurations were not high and flat as would be expected, at least when the ideal item pool was used in which difficulty and discrimination parameters were uncorrelated.

Gorman (1980) also examined the bias and information of scores produced by Owen's (1969, 1975) Bayesian testing procedure. These analyses were based on two "ideal" item pools with discriminations of $a = .8$ and 1.6, in which 101 items were rectangularly distributed in difficulty and both true and estimated item parameters were used. Gorman also studied the effect of applying a correction for regression (proposed by Urry, 1977) to θ estimates from Owen's testing procedure, designed to reduce bias in the estimates. His results showed substantial bias in the uncorrected θ estimates, with positive bias for θ levels below zero, negative bias for θ levels above zero, and higher levels of bias for the less discriminating items. His data also showed that Urry's correction was not entirely successful in eliminating the bias, since the corrected θ estimates for θ levels above zero resulted in positive bias. Since Gorman's study used an ideal, but finite, item pool, however, his results may be partially item pool dependent. In addition, Gorman's study did not attempt to determine the cause of the bias in the θ estimates but simply examined one possible approach to reducing it.

Purpose

The present study was designed to further investigate the nature of the bias and the information characteristics of Owen's (1969, 1975) Bayesian adaptive testing strategy and to examine possible causes of the bias. Factors investigated included (1) the effects of an accurate prior θ estimate, (2) the effects of item discrimination, and (3) the effects of fixed versus variable test length.

Method

Design

Monte carlo simulation of Owen's adaptive test was used. Unlike some previous simulation studies, but similar to Studies 1 to 3 in McBride (1977), the present studies did not use a prestructured item pool. Rather, the tests were simulated using a perfect and infinite item pool having any difficulty parameters required by the item selection process, with restrictions only on the item discriminations and pseudo-guessing parameters, c. By thus simulating an infinite item pool, the results of the sim-

ulation studies should reveal, within the limits of sampling error, the inherent properties of the Bayesian adaptive test, unaffected by the idiosyncrasies of a typical finite item pool.

Similarly, following the procedures of Study 4 in McBride (1977) in order to permit accurate description of the properties of the testing method as they vary with θ level, the simulated examinees (simulees) were not drawn randomly from a specified distribution; rather, a large number of examinees were simulated at each of a number of θ levels throughout the normally encountered range.

Examinees

For the purposes of monte carlo simulation, an examinee i was characterized by a numerical value, which is the actual trait level θ. In each of the eight data sets generated, there were 3,100 simulees, with 100 at each of 31 θ levels equally spaced in the interval −3.0 to 3.0. This range of the trait would include 99.99% of a population normally distributed on θ, with mean 0 and variance 1.

Test Items

For each separate item administration, an item was computer generated with the pseudo-guessing parameter, c, held constant at .20, simulating a five-alternative multiple-choice item. The item discrimination, a, was constant for each data set, with $a = .80$, 1.60, or 2.40 between data sets.

Following McBride (1977), the difficulty parameter, b, for each simulated item administration was determined by the current θ (the prior mean, M_{k-1}, of the estimated distribution of θ_i before administering the kth item) and by the constant item parameters a_g and b_g, according to the formula

$$b_g = M_{k-1} - \frac{1}{1.7a_g} \log \left[\frac{1 + (1 + 8c_g)^{1/2}}{2} \right]. \quad (1)$$

Equation 1 gives the item difficulty value having maximal information when $\theta_i = M_{k-1}$, and a_g and c_g are fixed (Birnbaum, 1968, p. 464). Since, in general, θ_i is unknown and the best available estimate is M_{k-1}, the item difficulty chosen is the one that is the most informative, given the current estimate of θ at any point in the adaptive test.

Item Response

The dichotomous (0,1) score of any simulee on any item is a probabilistic function of its status θ_i on the trait θ, of the item difficulty b_g, and of the parameters a_g and c_g. The probability $P'_g(\theta_i)$ of a correct response ($u_g = 1$) under the logistic model item characteristic curve is

$$P'_g(\theta_i) = \frac{c_g + (1 - c_g)}{1 + \exp[-1.7a_g(\theta_i - b_g)]} \quad . \quad (2)$$

In order to simulate item responses, each time an item administration took place, the quantity $P'_g(\theta_i)$ was compared with a pseudo-random number r_{gi} generated from a distribution uniform in the interval [0,1]. A score of $u_g = 1$ was assigned whenever $P'_g(\theta_i)$ equaled or exceeded r_{gi}; otherwise, a score of 0 was assigned.

Dependent Variables

For the simulated test of each individual i, the following were recorded:

k, the number of items administered;
M_k, the posterior mean after k items (i.e., $\hat{\theta}$); and
V_k, the posterior variance after k items (i.e., the variance of $\hat{\theta}$).

These values were averaged at each level of θ across the 100 simulees at that level, resulting in $\bar{\hat{\theta}}_i$, the mean of the θ estimates at each level of $\theta_i (i = 1, 2, \ldots, 31)$, and in $\sigma^2(\theta_i)$, the variance of $\hat{\theta}$ at each θ level. Bias was determined at each of the θ levels by

$$\text{Bias} = (\bar{\hat{\theta}}_i - \theta_i) \quad . \quad (3)$$

Information was computed from the formula

$$I(\theta_i) = \bar{\hat{\theta}}'^2_i / \sigma^2(\hat{\theta}_i) \quad , \quad (4)$$

where $\hat{\theta}'_i$ is the first derivative of the polynomial regression of $\hat{\theta}$ on θ.

Independent Variables

Eight data sets were analyzed for three levels of

item discrimination. The characteristics of the three studies and the data sets are summarized in Table 1.

Study I: Accurate prior θ estimate. This study was intended to provide "best case" data in order to serve as a benchmark against which other studies could be evaluated. The "best case" for the Bayesian adaptive test ought to be one involving a "perfect" item pool and accurate prior knowledge about examinees' trait levels. Accurate prior knowledge means that each examinee's trait level was known beforehand and was used as the mean of the Bayes prior distribution. Under these conditions, the only limitations on the information and accuracy of estimate of Owen's procedure are those imposed by the test length and by the discriminations and guessing parameters of the simulated test items. Holding those variables constant, any idiosyncrasies in the behavior of the test scores must be due either to the θ level estimation or item difficulty selection procedure.

Two separate and independent test administrations were simulated for each of the 3,100 simulees: in Data Set 1, all item discriminations were .80, and in Data Set 2, $a = 1.60$. For each simulee, the Bayes initial prior distribution was normal, with mean θ_i and variance 1.0. Thus, at the outset of testing, the initial estimate of each simulee's θ level was accurate. The adaptive test was allowed to run its normal course, reestimating θ_i after every item response and selecting the next item accordingly, until 20 items had been administered.

Study II: Constant prior θ estimate with fixed test length. Study II replicated the 20-item fixed test length and constant a values of .80 and 1.60 from Study I. To examine effects with more highly discriminating items, Data Set 5 used $a = 2.40$ for all items, whereas Data Sets 3 and 4 used items with $a = .80$ and 1.60 as in Study I. In contrast to Study I, the three data sets of Study II used the same initial normal prior distribution (mean = 0, variance = 1) for all simulees, regardless of actual θ level. In this study, then, a more typical use of the Bayesian adaptive testing strategy was simulated, that is, the application to individuals for whom no prior θ estimates were available prior to testing; consequently, a group prior θ distribution was used to select the first item to be administered. As in Study I, a fixed-length test of 20 items was administered to each simulee.

Study III: Constant prior θ estimate with variable test length. In Study III, as in Study II, the same initial normal (0,1) prior distribution was assumed for all simulees. The difference between the studies was in the test termination criterion. In

Table 1
Summary of the Independent Variables in the Three Studies

Study and Data Set	a	Prior Distribution		Termination Criterion	
		Mean	Variance	Posterior Variance	No. of Items
Study I					
1	.80	θ_i	1	-	20
2	1.60	θ_i	1	-	20
Study II					
3	.80	0	1	-	20
4	1.60	0	1	-	20
5	2.40	0	1	-	20
Study III					
6	.80	0	1	.10	30
7	1.60	0	1	.10	30
8	2.40	0	1	.10	30

Study III, testing was terminated for each simulee whenever the posterior variance V_k fell below .10. This value corresponds to the "standard error of estimate" criterion of .3162 specified by Urry (1974) to achieve a fidelity coefficient exceeding .95 in a normal (0,1) population of examinees. A maximum test length of 30 items was imposed, so that if the posterior variance criterion had not been reached within 30 items, testing was terminated. As for Study II, three levels of item discrimination—a = .80, 1.60, and 2.40—were studied in Data Sets 6, 7, and 8, respectively.

Results

Accurate Prior θ Estimate

Bias of the θ estimates for the two data sets of Study I is shown in Figure 1. As Figure 1 shows, there is virtually no bias in the θ estimates for Data Set 2 (a = 1.6), with a small amount of bias alternating between positive bias and negative bias for Data Set 1 (a = .8). The maximum amount of bias observed in the data is at θ = +3, where mean bias is −.10; a similar degree of bias is observed at θ = −1.8.

Figure 2 shows information curves for Data Sets 1 and 2. As the results show, the information for Data Set 1 is relatively flat throughout the θ range. The maximum information is at θ = −.5, with minimum information at θ = +.2. Information ranges between 7 and 11, with only minor variations across the θ range. The information for Data Set 2 is relatively flat, but not as flat as that for Data Set 1. There is a spike at θ = .8 with a secondary peak at θ = −2.8, and overall there is more variability between θ levels than for Data Set 1. In general, there is a slight concave trend to the information values for Data Set 2, with the exception of the spike at θ = .8. However, the general trend is a relatively flat information function for both data sets.

Constant Prior θ Estimate with Fixed Test Length

Figure 3 shows the bias in the θ estimates for the data sets of Study II at each of the three levels of item discrimination. For all three data sets there is a negative slope to the bias curve with low θ values being overestimated and higher θ values being underestimated. In addition, there are some substantial differences in the bias curves for the three levels of discrimination. Data Set 3 (a = .8)

Figure 1

Bias as a Function of θ for Data Sets 1 and 2

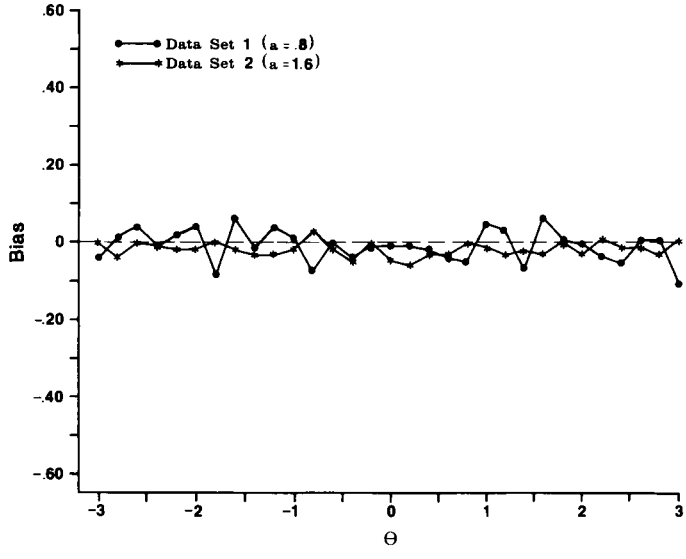
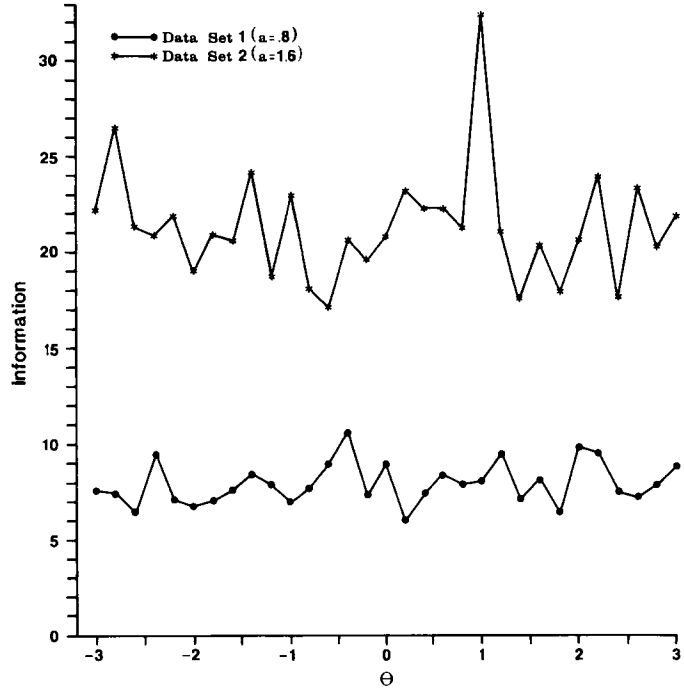


Figure 2
Information as a Function of θ for Data Sets 1 and 2



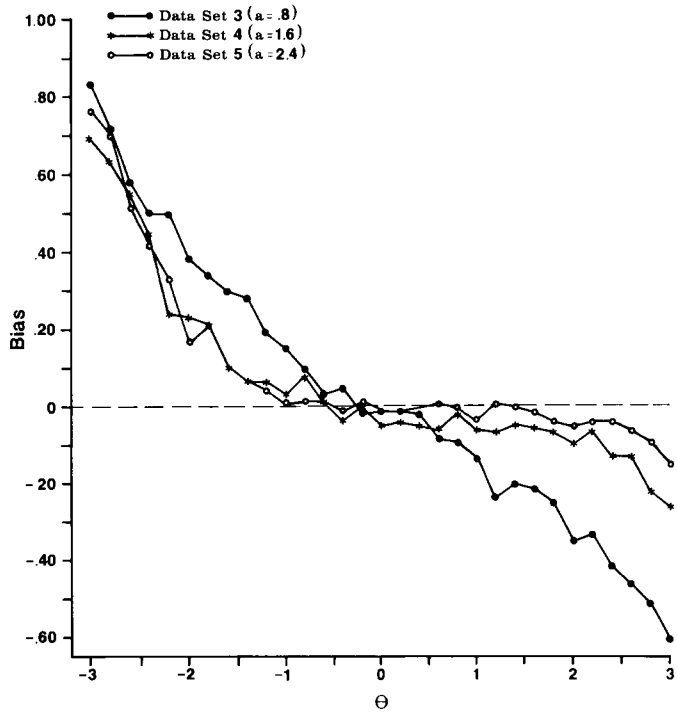
has the highest levels of bias of all three data sets. There is very severe bias for negative θ levels and severe bias in the opposite direction for positive θ levels. When item discriminations were increased in Data Set 4, there is only a slight drop in the positive bias for low θ levels and a more substantial drop in negative bias for the θ levels above the mean. Increasing the item discriminations to 2.4 in Data Set 5 resulted in virtually no change in bias for low θ levels, but it did result in a further decrease in bias for the positive θ levels with the range of unbiased θ estimates varying from approximately $\theta = -1$ to $\theta = +1.5$ in Data Set 5. As these results show, the effect of increasing item discrimination is to reduce bias somewhat, primarily for high θ levels. For low θ levels (< -2.0), substantial levels of bias (.20 or more) were observed for the highly discriminating items of Data Set 5.

Figure 4 shows test information curves for the three data sets of Study II. As Figure 4 shows, with the low discriminating items ($a = .8$) of Data Set 3, test information is relatively flat for θ levels above about $\theta = -1.5$, with a decrease in information below that level. As item discrimination is increased, the results for Data Set 4 show the information curve peaking with relatively lower information levels for $\theta > 1.6$ and $\theta < -1.5$ and becoming asymmetric. Finally, when the items of Data Set 5 ($a = 2.4$) were used, the information curve becomes even more peaked and more variable, with high levels of information generally in the range of $\theta = +1$ to -1 and with information dropping off extremely quickly beyond that range. For θ levels below -1, there is little difference in information when item discriminations are increased from $a = 1.6$ to $a = 2.4$. For θ levels below -1.8, levels of information are not increased by increasing item discriminations.

Constant Prior θ Estimate with Variable Test Length

Figure 5 shows bias functions for the three data

Figure 3
Bias as a Function of θ for Data Sets 3, 4, and 5

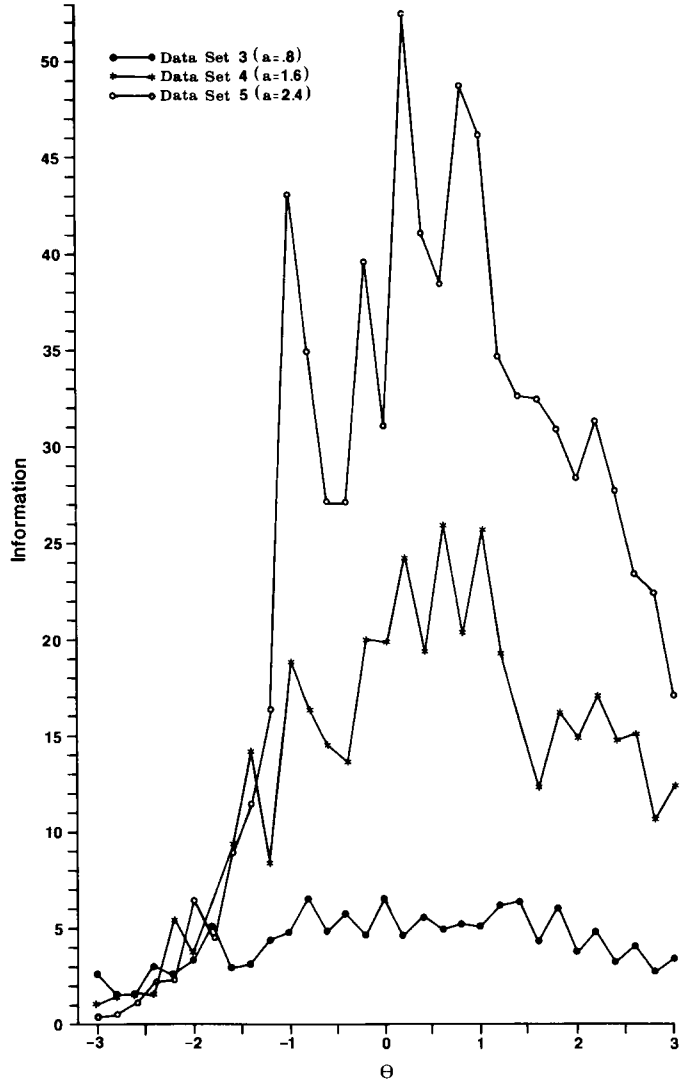


sets of Study III. As the results show, least bias for low θ levels was observed for Data Set 6 ($a = .8$), whereas the high θ levels obtained the highest degree of bias for that data set. As item discriminations increased, bias for low θ levels increased, while bias for the high θ levels decreased. Extremely high levels of bias were observed for Data Set 7 ($a = 1.6$) and Data Set 8 ($a = 2.4$) for θ levels less than $\theta = -2$.

Figure 6 shows test information functions for the variable-length conditions of Data Sets 6 through 8. The information function that most approximated the horizontal and equiprecise ideal was achieved by Data Set 6 ($a = .8$), which obtained relatively constant levels of information for θ values greater than $\theta = -1.5$. As item discrimination was increased, the level of information obtained for low θ levels decreased, while the level of information obtained for high θ levels remained similar. The result of increasing item discrimination was a general increase in peakedness and asymmetry of the test information functions.

Figure 7 shows the mean number of items administered for each of the θ levels for the data sets of Study III. As expected, more items were needed in Data Set 6, which had lower item discriminations, than in Data Sets 7 and 8. The results show that in Data Set 6, 30 items were generally not sufficient, on the average, for the adaptive test to achieve the specified level of posterior variance (.10) for most test lengths. The results also show that test length required was an increasing function of θ for Data Sets 7 and 8. Although, on the average, the posterior variance termination criterion of .10 was achieved with about 8.5 items for low θ values in Data Set 7, twice the number of items (17.0) were necessary to achieve the same posterior variance termination criterion (on the average) for $\theta = +3$. The same trend was observed for the more highly discriminating items of Data Set 8.

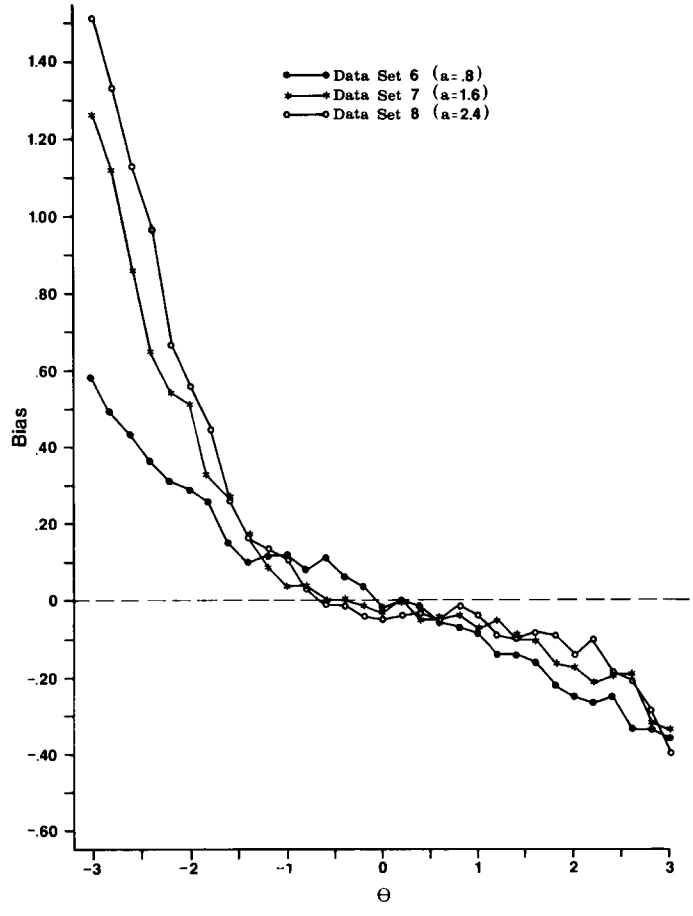
Figure 4
Information as a Function of θ for Data Sets 3, 4, and 5



Discussion and Conclusions

This study used a "perfect" item pool in order to evaluate the performance of Owen's (1969, 1975) Bayesian adaptive testing strategy under ideal conditions. The results show that in terms of achieving statistically unbiased measurement and measurements of equal precision throughout the θ range, Owen's adaptive testing strategy achieves these desirable goals only under the extremely unrealistic condition of an accurate prior θ estimate. In a realistic testing situation, the examinee's trait level is not known beforehand; otherwise, testing would not be necessary. Thus, the data of Study I serve only as an unrealistic baseline condition to which results of other more realistic testing conditions can be compared. Even under the unrealistic conditions of Study I, however, there was a tendency for in-

Figure 5
Bias as a Function of θ for Data Sets 6, 7, and 8

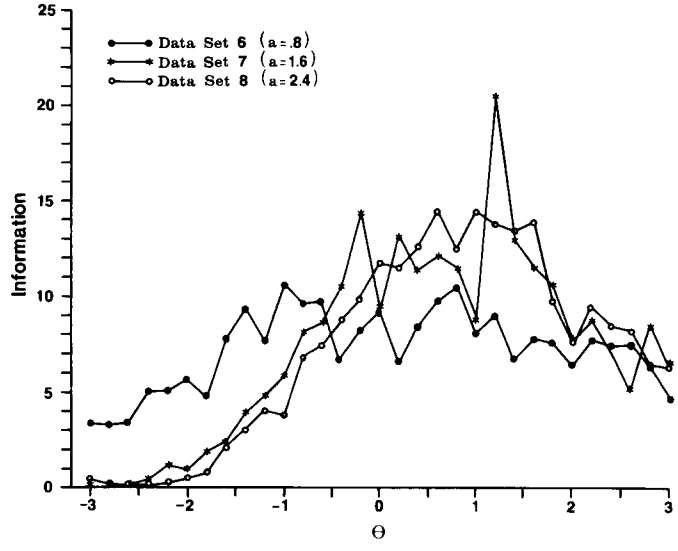


creasing item discrimination to result in increasing variability in levels of information as a function of θ.

Studies II and III evaluated Owen's Bayesian testing strategy under the more realistic testing conditions of a constant prior θ estimate, with both fixed and variable test length. The results of Studies II and III show that this adaptive testing strategy does not achieve unbiased measurement or measurements of equal precision when a constant prior θ estimate is used for all examinees, regardless of whether test length is fixed or variable. The results show an interaction of the termination criterion with the performance of the adaptive testing strategy, both in terms of bias and information.

When a constant test length is used, increasing item discrimination results in decreased bias, with a more substantial decrease in bias for high θ levels. When variable termination is used, increasing item discrimination results in only slightly decreased bias for high θ levels but in *increased* bias for low θ levels, with extremely high levels of bias for very low θ levels. In terms of information, the flattest information curves were observed for both termination criteria with the least discriminating items. As item discrimination was increased, in both cases the information curve became more peaked and asymmetric, with a greater degree of asymmetry observed for the variable-length testing condition. Results also showed that different mean numbers

Figure 6
Information as a Function of θ for Data Sets 6, 7, and 8



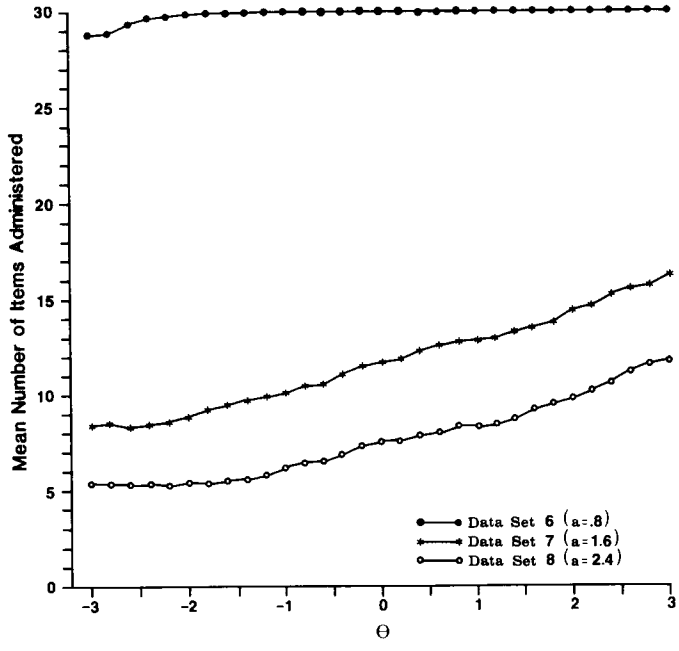
of items were necessary to achieve a fixed posterior variance termination criterion at different levels of θ. With moderately and highly discriminating items (a = 1.6 and a = 2.4), twice the number of items were necessary, on the average, for high θ levels to reach a posterior variance termination criterion of .10 than for low θ levels.

Because this study used a perfect item pool in which items of a specified discrimination were available at any level of difficulty, the results observed in these studies cannot be attributed to deficiencies in the item pool, as might be the case for the results reported by Gorman (1980). Rather, these results are attributable to the effect of the constant prior θ estimate, as is shown by the comparison of results between Studies II and III and those of Study I. Although the effects of Urry's (1977) correction for regression was not explicitly examined in these studies, it is unlikely that it would have the desired effects under both the fixed-length and variable-length test condition, since, as indicated, there was interaction of observed bias with the termination criterion.

Since an accurate prior θ estimate resulted in no bias and a constant prior θ estimate resulted in substantial bias, it can be assumed that differential priors of less than perfect accuracy will result in some degree of bias in the θ estimates. However, an inaccurate prior for an individual will have the same effect on bias as will a constant prior where the prior is distant from the examinee's true θ. Thus, an inaccurate prior above an examinee's θ will result in a $\hat{\theta}$ with positive bias, whereas a prior below an examinee's θ will result in negative bias. Furthermore, in many cases prior information may not be available at the individual level. In both these situations, Bayesian adaptive testing will result in measurements with less than optimal characteristics.

Wood (1971) observed bias in θ estimates from Owen's Bayesian adaptive test. He attributed the bias to the effect of the c parameter on the θ estimates, indicating that it affects the mean and variance of the Bayesian θ estimates as well as influencing the choice of the item selected (pp.134–135). He suggested that overcorrecting for guessing, as may be the case when a constant c is used (as was done in the present study), will result in underestimation of θ. This might explain the results observed here for θ levels above the prior, but it would not explain the overestimation of θ observed for θs below the prior. Wood (1971, pp.151–155)

Figure 7
Mean Number of Items Administered as a Function of θ for Data Sets 6, 7, and 8



also suggested that the bias he observed in the Bayesian θ estimates might result from an uneven distribution of the item parameters through the effect of item discriminations on the movement of the θ estimates. In the current studies, however, item discriminations were rectangularly distributed (and item difficulties were exactly those required by the item selection procedure), yet severe bias was still observed. Thus, the present results indicate that the bias in the Bayesian θ estimates derives from the use of an incorrect Bayesian prior θ estimate and not from the effects of either guessing or the distribution of the *a* and/or *b* parameters.

Although a major purpose of adaptive testing is to provide measurements with equal precision/information at all levels of the trait continuum (Weiss, 1982), results of these analyses show that under the realistic conditions of a constant prior θ estimate, Owen's (1969, 1975) Bayesian adaptive testing strategy does not achieve this desirable goal. Since the test information curves utilize some of the same data from which the bias curves were computed, the results for information are in a sense a consequence of the bias in the θ estimates. The data from these three studies show that the bias results from use of a constant prior θ estimate. Further research will be necessary to determine whether and to what degree the use of variable prior θ estimates will affect the performance of Owen's adaptive testing strategy in terms of reducing the bias and, consequently, improving the equiprecision of its trait level estimates.

By contrast with the present results, either Owen's Bayesian item selection procedure or maximum information item selection in conjunction with maximum likelihood θ estimation result in unbiased θ estimates and equiprecise measurements, even when a constant θ level is used to select the first item for administration (e.g., Weiss, 1982). Thus, maximum likelihood θ estimation may be preferable to the Bayesian approach when no differential prior information is available for an examinee, or when the prior information that is available might be inaccurate.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley, 397–472.

Gorman, S. (1980). *A comparative evaluation of two Bayesian adaptive ability estimation procedures with a conventional test strategy*. Unpublished doctoral dissertation, Catholic University of America, Washington DC.

Jensema, C. J. (1972). An application of latent trait mental test theory (Doctoral dissertation, University of Washington, 1972). *Dissertation Abstracts International, 24,* 633. (University Microfilms No. 72–20,871).

Jensema, C. J. (1974). The validity of Bayesian tailored testing. *Educational and Psychological Measurement, 34,* 757–766.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989–1020.

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

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

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

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

Urry, V. W. (1971, April). *Individualized testing by Bayesian estimation* (Research Bulletin 0171–177). Seattle: University of Washington, Bureau of Testing.

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

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14,* 181–196.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–491.

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

Acknowledgments

This research was supported by Contract N00014–79–C–0172, NR150–433, from the Office of Naval Research, with additional funding from the Army Research Institute, Air Force Office of Scientific Research and the Air Force Human Resources Laboratory.

Author's Address

Send requests for reprints or further information to David J. Weiss, Department of Psychology, N660 Elliott Hall, University of Minnesota, Minneapolis MN 55455, U.S.A.