A Comparison of the Accuracy of Bayesian Adaptive and Static Tests Using a Correction for Regression

Steven Gorman
Department of the Navy

The vast changes in computer technology have made a strong impact upon the field of ability measurement. The increased capabilities and decreased costs of computer use have opened the door to application of latent trait theory. Two Bayesian procedures for ability estimation have become popular--the Bayes modal procedure (Samejima, 1969) and the Owen (1975) algorithm. Both Bayesian procedures use a prespecified distribution, usually the Gaussian normal distribution, as the prior variance of ability. The item characteristic curve (ICC; also called the item response function) is employed as the likelihood function. The product of the prior distribution and the likelihood function is the posterior distribution of ability. These two procedures can be used in either conventional or adaptive mode.

McBride and Weiss (1976) have studied Owen´s Bayesian adaptive procedure and have determined that with this procedure, ability estimates regress toward the mean. That is, high-ability examinees tend to achieve lower ability estimates, and low-ability examinees tend to have higher ability estimates. Urry (1977) has suggested a correction, namely, dividing the Bayesian regressed ability estimate by the test reliability. A second, potentially more serious, problem is the reliance upon accurate 3-parameter logistic item parameters. Urry (1976) developed OGIVIA3, a computer program to estimate these item parameters. The effectiveness of this estimation procedure for use in the Owen algorithm was reviewed by Gugel, Schmidt, and Urry (1976). OGIVIA3 has been revised (Croll & Urry, in prep.) and has been renamed ANCILLES.

The purpose of the present paper is to evaluate the effectiveness of two Bayesian ability estimation procedures with a correction for regression using known and estimated parameters. Specifically, the studies simulated the Owen algorithm and Bayes modal testing methods in both adaptive and static mode with a correction for regression using known parameters and the parameters estimated using ANCILLES.

Study 1: An Analysis of the Verbal Scholastic Aptitude Test

Background and Purpose

Lord (1968) applied the 3-parameter logistic model developed by Birnbaum (1968) to the Verbal Scholastic Aptitude Test (VSAT). Until Lord´s article, little research had been conducted using Birnbaum's model. However, since this article, with the exception of a few articles involving the maximum likelihood procedure (Bejar, Weiss, & Gialluca, 1977; Kolakowski & Bock, 1970, 1972; Wood, Wingersky, & Lord, 1976), the overwhelming majority of latent trait research has applied the work of Birnbaum to adaptive tests and not to conventional tests. Samejima (1968) detailed the mechanics of a Bayes ability estimator based on a response pattern of test items. She proved that with an assumed normal distribution of ability as a prior distribution, and using the ICC as a likelihood function, the mode of the posterior distribution will provide an absolute maximum, which can be used as an ability estimate. Urry (1976) incorporated the Bayes modal procedure in the second stage of his item parameter estimation program. Owen (1975) developed a Bayesian procedure for estimating ability; however, this procedure was developed for the adaptive mode. Bejar and Weiss (1979) programmed the Owen algorithm for scoring static tests, but no data on its effectiveness were made available.

The purpose of this study was to investigate the efficiency of the Bayes modal and Owen's Bayesian ability estimation procedures relative to a conventional rights-only scoring. In particular, the issues investigated are (1) conditional bias, (2) conditional accuracy, and (3) precision of test scores.

Design of the Study

Artificial data were generated according to the 3-parameter logistic model:

$$P_i(\theta) = c_i + (1 - c_i) [1 + \exp(-1.7a_i(\theta - b_i))]^{-1} \qquad [1]$$

using the LVGEN program developed by Urry (1971). This program provided vectors of responses, correct (1) or incorrect (0), for the simulated examinees (sims). The test items used had the parameters of the first 80 VSAT items reported in Lord (1968).

For the purpose of this study, it was assumed that the item parameters reported in Lord's study were the actual parameters and not estimated, as they actually were. The 80 item parameters were administered to 2,000 sims from a normal distribution (mean 0, variance 1) generated by the LLRANDOM Computer Program (Learmonth & Lewis, 1973) in conjunction with the LVGEN program. The resulting vectors of simulated binary responses were analyzed by the ANCILLES Program; estimates of the 80 "known" VSAT parameters were the resultant output. This allowed a comparison of the robustness of the Bayesian ability estimation programs to inaccuracy in the item parameter estimates. An additional 2,000 normally distributed sims were administered the VSAT items. This permitted computation of the correlation of known ability with the various ability estimates and the mean and variance of raw scores so that a Z-transformation could be computed. This allowed comparison of a simpler scoring procedure based on classical test theory with the two scoring procedures based on latent trait theory.

Five conditions of scoring the same item responses were examined: (1) Bayes modal ability estimates based on known item parameters, (2) Bayes modal ability estimates based on estimated item parameters, (3) Owen's Bayesian ability estimates based on known item parameters, (4) Owen's Bayesian ability estimates based on estimated item parameters, and (5) ability estimates based on raw score to Z-score transformations.

To properly address the evaluation mentioned above required examination of the test score characteristics as a function of ability level. Therefore, the ability distribution consisted of 100 sims at each of 11 equally spaced values in the interval $-2.5 \leq b \leq +2.5$.

For each of the five simulated test administrations, conditional bias, conditional accuracy, and conditional precision were estimated from the 100 observations at each ability level (θ_e).

Conditional bias. This statistic provided an indicator of the magnitude and direction of the error between true ability and ability estimated by each of the scoring procedures at various levels of the trait continuum where

$$\text{bias} = b_e | \theta_e = \bar{\hat{\theta}}_e - \theta_e \text{ ,} \qquad [2]$$

where

b_e = average bias for each of 11 values of on the trait continuum,

θ_e = true ability of examinees for each value, and

$\hat{\theta}_e$ = average ability estimates for each value.

Conditional accuracy. The accuracy of the test scores was provided by the root mean square error computed for the 11 values using the formula

$$e_i | \theta = \left[\sum_{i=1}^{n} (\theta_e - \hat{\theta}_e)^2 \; n^{-1} \right]^{\frac{1}{2}} \qquad [3]$$

where

$e_i | \theta$ = root mean square error conditional upon ability level,

n = 100,

$\bar{\theta}$ = known ability level, and

$\hat{\theta}_e$ = the ability estimate.

Conditional precision. This statistic was provided by the test score information function. The information generated by a score about a given ability level can be compared to the precision of measurement at that point. Samejima (1977) stated that the inverse of the square of information can be considered as the standard error of measurement when number of items and test information are sufficiently large. Birnbaum (1968) provides a formula for information:

$$I_e(\theta') = \left[\frac{\partial}{\partial\theta} \; \frac{E(\hat{\theta}_e | \theta}{\hat{\sigma}_{\hat{\theta}_e} | \theta} \right]^2 \qquad [4]$$

where I_e (θ') is the information about θ provided by score $\underline{x}$. Sim scores were calculated at each of 11 equally spaced ability levels $-2.5 \leq x \leq +2.5$; these test score means were used to estimate the slope by fitting a curve through three consecutive values. Because test score means were required on either side of the information point, information values could not be computed for the ends of the continuum (−2.5, +2.5).

Results

Estimation bias. The comparisons between the two Bayesian procedures for scoring static tests using estimated parameters and the raw score to Z-score transformation are in Figure 1. The figure shows that the absolute value of bias for the Z-score was much greater than for the two Bayesian procedures at ability level −2.5. The absolute value of Bayesian score bias tended to be equal to or lower than that of the Z-score along the entire trait continuum. Of the two Bayesian procedures, the Bayes modal bias was greater at upper trait levels.

Figure 1
Bias of Three Scoring Procedures, Using
Estimated Item Parameters in a Static Test

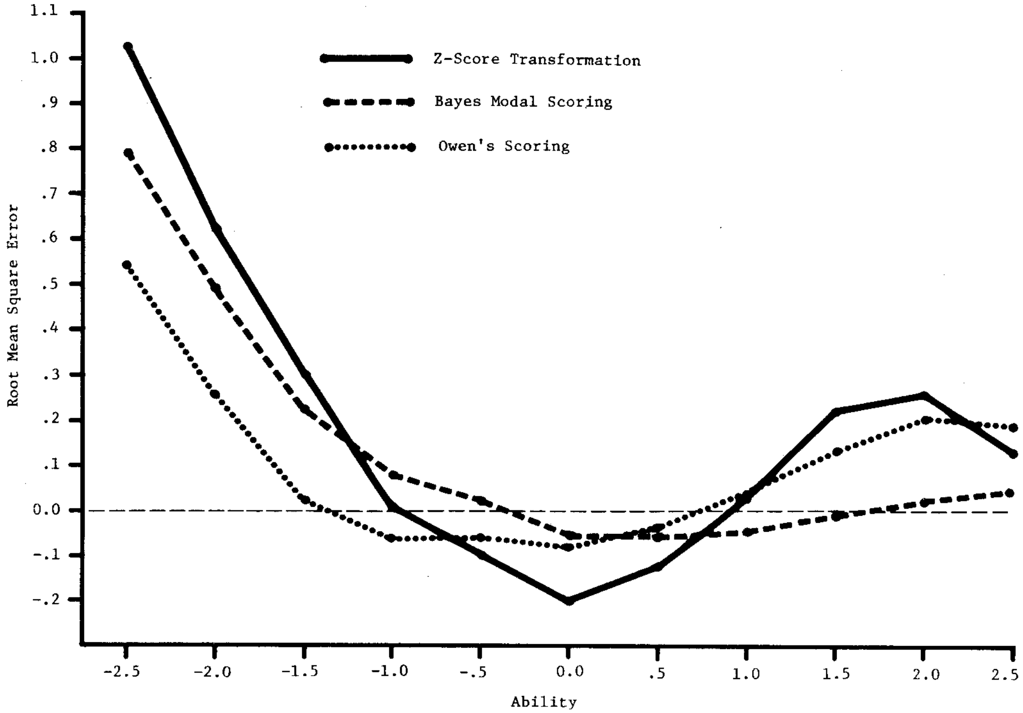
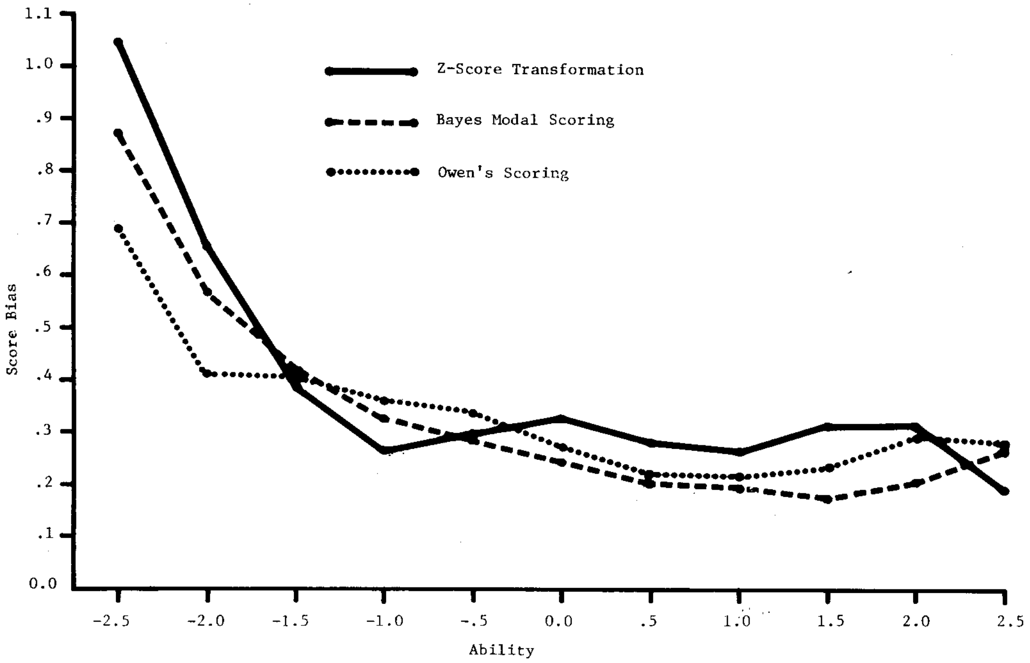


Table 1 shows the bias values of the two Bayesian procedures under conditions of known and estimated parameters, as well as the conventional Z-score method. The Bayes modal scores using known parameters still suffered to some degree from the regression to the mean effect, although deviations from zero were mostly lower than the bias from either estimated Bayes or Z-score methods. Improvements to the estimation of item parameters could decrease the bias of the two Bayesian static procedures significantly.

Table 1
Bias of Conventional Z-Score Method and Two Bayesian Scoring Methods Using Estimated and Known Parameters in a Static Test

		Parameters			
		Estimated		Known	
Ability Level	Z-Score	Bayes Modal	Owen´s	Bayes Modal	Owen´s
-2.5	1.027	.789	.541	.482	.366
-2.0	.621	.488	.253	.260	.059
-1.5	.300	.221	.022	.087	-.143
-1.0	.010	.077	-.062	.065	-.142
-0.5	-.097	.022	-.058	.051	-.083
0.0	-.200	-.056	-.081	-.019	-.064
0.5	-.126	-.058	-.035	-.023	-.008
1.0	.031	-.048	.037	-.027	.046
1.5	.220	-.013	.132	-.030	.104
2.0	.260	.023	.206	-.044	.138
2.5	.130	.044	.188	-.060	.109

Figure 2
Root Mean Square Error of Three Scoring Methods Using Estimated Item Parameters in a Static Test



Conditional accuracy. Figure 2 displays the root mean square error (RMSE) of ability estimation for the two Bayesian algorithms using estimated parameters and the Z-score method. All three methods followed the same trend of having high RMSE values at the low-ability levels and diminishing asymptotically to a value of about .2 at the trait level +.5. This phenomenon appeared to be a function of the test itself, with its emphasis on more precise measurement at the higher ability levels. The conventional scoring procedure tended to have the highest inaccuracy, with two exceptions (ability levels -1.0 and +2.5). Table 2 lists the RMSE values for the two Bayesian methods using known and estimated parameters.

Table 2
Root Mean Square Error of the Z-Score Method
and Two Bayesian Scoring Methods Using
Estimated and Known Parameters in a Static Test

		Parameters			
		Estimated		Known	
Ability Level	Z-Score	Bayes Modal	Owen´s	Bayes Modal	Owen´s
-2.5	1.048	.875	.686	.703	.537
-2.0	.652	.567	.412	.486	.365
-1.5	.386	.418	.405	.482	.463
-1.0	.263	.325	.361	.370	.425
-0.5	.296	.284	.338	.300	.382
0.0	.328	.243	.273	.241	.288
0.5	.281	.203	.221	.191	.213
1.0	.264	.195	.215	.185	.212
1.5	.313	.176	.233	.160	.204
2.0	.314	.205	.293	.188	.239
2.5	.191	.266	.280	.252	.231

Conditional precision. The test score information values at the nine ability levels, -2.0 to +2.0, for the two Bayesian scoring methods using estimated parameters and the conventional scoring procedure, are in Figure 3; numerical values are in Table 3. The data in Table 3 coincide with two trends of the earlier study (Lord, 1968, p. 998) on the VSAT. First, the data in Table 3 (as well as in Table 2) illustrate the more precise measurement on the VSAT at upper ability levels. Second, the data show that significant increases in precision can be gained by using the Bayesian scoring procedures.

The original study weighted items based on the logistic model and found this procedure provided greater information than conventional scoring. The average score information value for conventional scoring was 12.195; the average for the Owen scoring was 13.800 and was 14.120 for the Bayes modal scoring, with estimated item parameters used in the scoring procedures. Slightly higher averages (13.960 for the Owen and 14.503 for the Bayes modal scoring) occurred when the known item parameters were available.

Fidelity. Fidelity coefficients, the correlations of the known ability of 2,000 sims from a normal population with their estimated abilities, were computed from the various test scoring methods and are in Table 4. Although the increase in the correlation is only roughly .02 for the two Bayesian methods over

Figure 3
Test Score Information of Three Scoring Methods, Using Estimated Item Parameters in a Static Test

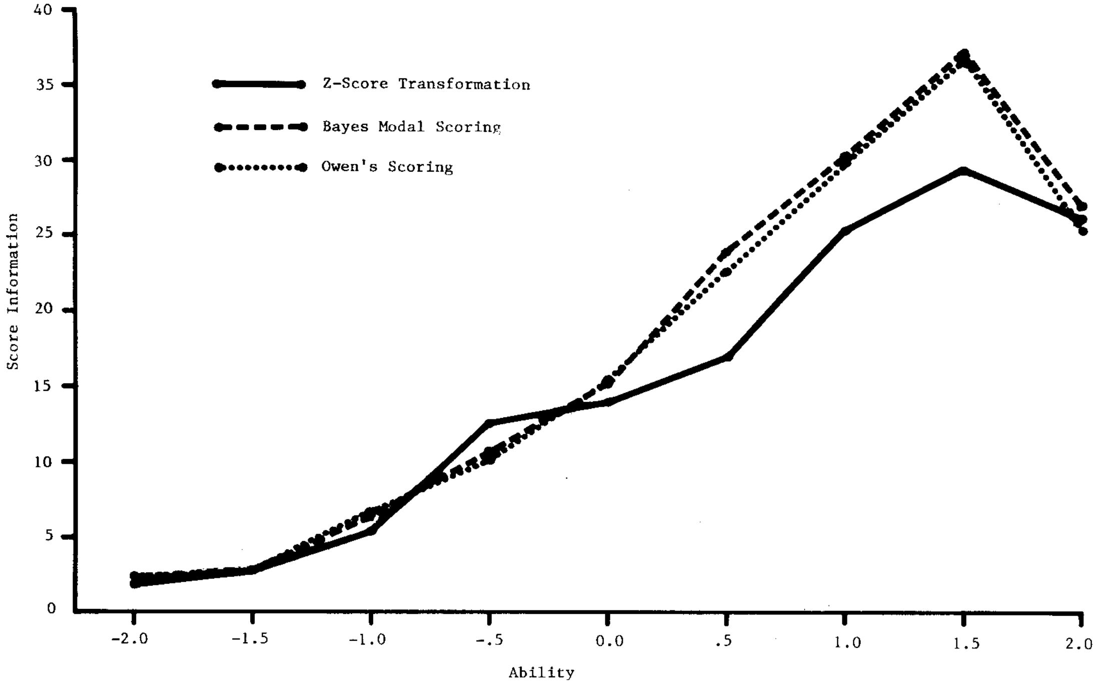


Table 3
Test Score Information of Conventional Z-Score Method and Two Bayesian Scoring Methods Using Estimated and Known Parameters in a Static Test

		Parameters			
		Estimated		Known	
Ability Level	Z-Score	Bayes Modal	Owen´s	Bayes Modal	Owen´s
-2.0	1.910	2.285	2.215	2.185	1.876
-1.5	2.775	2.713	2.817	2.796	3.177
-1.0	5.316	6.343	6.640	6.930	6.963
-0.5	12.545	10.645	10.189	9.986	9.407
0.0	13.963	15.200	15.484	14.807	14.725
0.5	16.876	23.884	22.631	26.225	24.184
1.0	25.307	30.190	29.836	29.441	28.582
1.5	29.386	37.179	36.691	38.875	38.288
2.0	26.066	26.880	25.294	28.293	26.353

the conventional method, at this high level (.94 to .96) the result is highly significant ($\underline{p} \leq .0001$). The fidelity coefficient of the 80-item VSAT test scored with either Bayesian method is comparable (via the Spearman-Brown proph-

Table 4
Correlation of Known Ability with Ability Estimates for a Conventional Z-Score Scoring Method and Two Bayesian Scoring Methods Using Known and Estimated Parameters in a Static Test

Scoring Method	r
Conventional Z-Score Transformation	.941
Bayes Modal	
Estimated Parameters	.959*
Known Parameters	.960*
Owen´s Bayesian	
Estimated Parameters	.958*
Known Parameters	.958*

*Values significantly different from conventional Z-score transformation r at $p \leq .0001$.

ecy formula) to a fidelity coefficient of a 120-item test scored conventionally. Also of interest is the fact that the fidelity coefficients computed using either Bayesian procedure with known item parameters were not significantly different from the fidelity coefficients computed from Bayesian scoring with estimated item parameters. This attests to the robustness of the Bayesian scoring procedure to errors in item parameter estimation.

Conclusions. It is apparent that improvements in the measurement of examinees on conventional tests can be realized by the use of mathematical scoring procedures that are based upon latent trait theory. Bias seems to be diminished, and test score accuracy and precision are improved with these two Bayesian scoring procedures, compared to the conventional scoring method.

Study 2: An Analysis of the Effect of the Correction for Regression and Parameter Estimation Errors Upon Two Bayesian Adaptive Testing Procedures

Purpose

The present study simulated an adaptive test using both Owen´s Bayesian procedure and the Bayes modal procedure. The research attempted to determine the effect of item parameter estimation errors upon the test characteristics as a function of ability level. In addition, this study investigated the effect of a correction for regression applied to the ability estimates obtained using the Owen algorithm. The Bayes modal procedure already incorporates this regression correction.

Owen´s Bayesian Procedure and the Correction for Regression

The Bayesian adaptive ability estimation procedure has been well documented elsewhere (McBride & Weiss, 1976; Owen, 1975) and will not be reported here. However, to understand the correction, a brief conceptual description is in order. The procedure assumes a normal distribution of the ability estimates with mean 0 and variance 1. The item bank is then scanned to identify the item that will minimize the expectation of the posterior variance of the distribution if administered. That item is then administered, and a new ability estimate (mean of posterior distribution) and variance about that estimate are computed. The ability estimate is then used as the prior mean, and an item is again selected to minimize the expected value of the variance of the posterior distribution. This procedure is repeated iteratively.

A correction for regression is applied to the final ability estimate. The correction consists of dividing the final ability estimate by what Urry (1977) refers to as the test reliability. This reliability is 1.0 minus the Bayesian posterior variance, and this value obviously will differ for each individualized test. Urry believes that more accurate measurement is attained by terminating adaptive tests based on a fixed posterior variance, rather than a fixed number of items. However, Urry (1977) concedes that this correction should be effective for both fixed and variable-length tests. This study investigates the fixed-length test only.

Bayes Modal Adaptive Procedure

The Bayes modal adaptive ability estimation procedure developed for this study consisted of two algorithms--one to estimate ability and one to select appropriate items to be administered. The ability estimation algorithm was based on the Bayesian scoring procedure developed by Samejima, using the item response function and an assumption of a normal distribution of ability. Urry (1976) uses this procedure in the second iterative stage of his item parameter estimation procedure. The item selection procedure chooses that item which provides the highest level of item information for the current ability estimate. The item response function for all administered items is computed. The product of all item response functions and the assumed normal density function is the posterior distribution; the mode of this distribution is the ability level estimate. This value is then unregressed using the same correction as stated earlier. However, unlike the Owen procedure, the corrected estimate is then used as the starting point for the next iteration of item selection.

Design of the Study

Two "ideal" banks were generated, each consisting of 101 items at equal increments of $\underline{b}$ = .05 over the range $-2.5 \leq \underline{b} \leq +2.5$. One bank used items whose item discriminations were set at $\underline{a}$ = 1.6; the other, at $\underline{a}$ = .8. The item parameters were estimated by the ANCILLES program on a group of 50 items based on the responses of 2,000 sims. The procedures differed from Study 1 in that the items were scrambled with the parameters from item banks of another study (Gorman, in prep.). The analysis was based upon three test characteristics as a function of ability level--bias, accuracy, and precison--as documented in Study 1.

Results

Conditional bias. Figure 4 displays the score bias from the 25-item adaptive test employing the Owen algorithm, with and without the correction for regression, and the Bayes modal procedure. The three lines represent the bias in the adaptive procedures using the item bank with item discriminations of a = 1.6, based on estimated parameters. The Owen procedure with the correction provided the least bias.

Figure 4
Effect of Regression Correction Upon Bias of
25-Item Bayes Modal and Owen Adaptive Tests
(a = 1.6) with Estimated Parameters

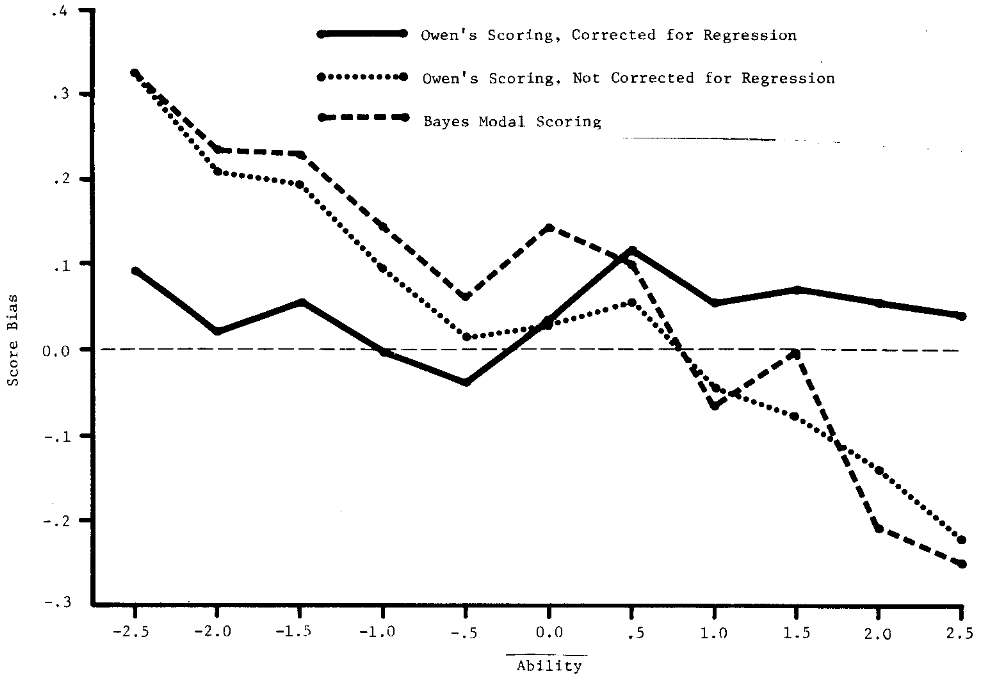


Table 5 shows the effect of regression upon the ability estimates from the Owen procedure using known and estimated parameters. An interesting result is that the regression phenomenon was more prevalent when known parameters were used in the Owen scoring with a correction than with estimated parameters using the same correction. This may be due to sampling errors in parameter estimation working in the preferred direction on this criterion.

Using the less discriminating item bank (a = .8), the regression was more extreme, but the correction using estimated parameters again adequately compensated. The regression correction was less effective when using known parameters.

The Bayes modal adaptive test did not fare as well as the Owen adaptive test. This can be seen in Table 6, which lists the bias for the two item banks under conditions of known and estimated parameters. With known parameters, the bias was tolerable with the better item bank. The bias under the other three conditions was significantly greater.

Conditional accuracy. Table 7 shows the effect of the regression correc-

Table 5
Effect of Regression Correction Upon Bias of the 25-Item Owen's Adaptive Test with Estimated and Known Parameters for Two Item Banks

Item Bank and Ability Level	Parameters			
	Estimated		Known	
	Corrected	Uncorrected	Corrected	Uncorrected
a=.8 Item Bank				
-2.5	.063	.468	-.055	.416
-2.0	.021	.351	-.167	.237
-1.5	-.036	.222	-.127	.179
-1.0	-.029	.146	-.124	.091
-0.5	-.053	.042	-.067	.043
0.0	.008	.007	.008	.006
0.5	.049	-.047	.102	-.019
1.0	.011	-.165	.075	-.143
1.5	-.027	-.283	.154	-.186
2.0	.046	-.306	.268	-.205
2.5	.036	-.403	.275	-.314
a=1.6 Item Bank				
-2.5	.091	.326	-.008	.242
-2.0	.019	.207	-.077	.125
-1.5	.055	.193	.001	.151
-1.0	-.004	.093	-.046	.061
-0.5	-.040	.013	-.068	-.009
0.0	.031	.028	.007	.006
0.5	.115	.055	.071	.010
1.0	.054	-.046	.064	-.050
1.5	.071	-.077	.117	-.059
2.0	.054	-.140	.161	-.077
2.5	.041	-.221	.153	-.150

tion upon the root mean square error (RMSE) of the 25-item Owen adaptive test. The average RMSE value for the Owen ability estimates using known item parameters without the correction was .225; using estimated item parameters with the correction, .233; using known item parameters with the correcton, .241; and using estimated item parameters without the correction, .225.

The a = .8 item bank followed this same trend, only to a greater degree, with the exception that the highest average RMSE value was with the Owen procedure using known item parameters and corrected for regression. This result is counter to the expected result. The reason for this may again be due to errors in item parameter estimation favorable to the Owen procedure. Another trend for both item banks was that the RMSE values were lowest about the mean and increased in magnitude as a function of distance from the mean.

Table 8 lists the RMSE for the Bayes modal adaptive test. On the item bank with a = .8 using estimated parameters, the conditional accuracy was poorer than

Table 6
Bias of the 25-Item Bayes Modal Adaptive Test
Using Estimated and Known Parameters,
with Two Item Banks

	Item Bank			
	a = .8		a = 1.6	
Ability Level	Estimated Parameters	Known Parameters	Estimated Parameters	Known Parameters
-2.5	.575	.213	.324	.115
-2.0	.382	.133	.232	.085
-1.5	.262	.101	.228	.156
-1.0	.094	.027	.143	.097
-0.5	.049	.035	.058	.047
0.0	.066	.043	.143	.059
0.5	.087	.084	.099	.066
1.0	-.049	-.013	-.067	-.002
1.5	-.067	-.101	-.002	-.038
2.0	-.195	-.092	-.208	-.058
2.5	-.343	-.080	-.251	-.038

Table 7
Effect of Regression Correction Upon Root Mean Square Error
of the 25-Item Owen Adaptive Test with Estimated and
Known Parameters for Two Item Banks

Item Bank	Parameters			
and	Estimated		Known	
Ability Level	Corrected	Uncorrected	Corrected	Uncorrected
a=.8 Item Bank				
-2.5	.370	.556	.413	.532
-2.0	.423	.497	.454	.417
-1.5	.404	-403	.384	.347
-1.0	.432	.388	.433	.349
-0.5	.368	.305	.387	.310
0.0	.352	.291	.390	.313
0.5	.447	.370	.420	.325
1.0	.401	.372	.445	.376
1.5	.351	.404	.416	.357
2.0	.339	.416	.531	.411
2.5	.395	.512	.541	.475
a=1.6 Item Bank				
-2.5	.2907	.4054	.2440	.3181
-2.0	.2411	.2973	.2481	.2434
-1.5	.2629	.3020	.2202	.2496
-1.0	.1864	.1926	.2088	.1927
-0.5	.2223	.1980	.2353	.2025
0.0	.2297	.2072	.2133	.1906
0.5	.2351	.1945	.2386	.2034
1.0	.2142	.1953	.2399	.2118
1.5	.2069	.1923	.2512	.2060
2.0	.2333	.2452	.2670	.2033
2.5	.2487	.2988	.2861	.2558

Table 8
Root Mean Square Error of the 25-Item Bayes Modal Adaptive Test Using Estimated and Known Parameters, with Two Item Banks

	Item Bank			
	a = .8		a = 1.6	
Ability Level	Estimated Parameters	Known Parameters	Estimated Parameters	Known Parameters
-2.5	.783	.445	.589	.393
-2.0	.593	.431	.393	.327
-1.5	.428	.373	.388	.279
-1.0	.366	.386	.290	.245
-0.5	.411	.384	.261	.264
0.0	.412	.351	.339	.290
0.5	.426	.414	.268	.233
1.0	.321	.327	.234	.228
1.5	.328	.395	.172	.215
2.0	.320	.368	.269	.206
2.5	.447	.340	.364	.210

Table 9
Test Score Information of Two 25-Item Bayesian Tests, Using Known and Estimated Parameters, with Two Item Banks

	Item Bank			
	a = .8		a = 1.6	
Adaptive Test and Ability Level	Known Parameters	Estimated Parameters	Known Parameters	Estimated Parameters
Owen´s Bayesian				
-2.0	2.591	4.738	15.776	17.847
-1.5	5.519	8.359	14.786	22.501
-1.0	5.311	6.475	23.878	21.238
-0.5	7.975	8.726	21.079	21.571
0.0	9.808	9.009	25.752	28.516
0.5	5.181	6.974	26.434	21.809
1.0	5.425	6.021	22.470	21.041
1.5	8.975	9.519	26.369	24.226
2.0	9.863	5.897	18.315	23.473
Bayes Modal				
-2.0	1.325	4.210	5.067	9.898
-1.5	2.850	5.822	5.435	13.496
-1.0	4.476	5.689	8.919	13.241
-0.5	4.859	6.425	11.277	11.670
0.0	6.347	8.906	10.608	12.359
0.5	4.982	5.695	17.014	18.389
1.0	7.565	6.504	24.089	15.992
1.5	6.621	5.594	21.752	19.452
2.0	5.142	7.703	8.609	23.604

with the Owen procedure. On the other hand, with the same item bank using known parameters, accuracy was greater with the Bayes modal procedure. With the better item bank, the Owen procedure was superior to the Bayes modal on this criterion.

Conditional precision. Table 9 lists values of score information for 25-item tests with both Bayesian adaptive methods and two item banks. The item parameter estimation errors rearranged the test score distribution and, hence, its information. The Owen procedure provided more information about the mean and dropped off somewhat at the extremes. The Bayes modal procedure provided considerably less information; hence, the standard error of measurement was larger at all ability levels.

Conclusions

The correction for regression effectively diminished the regression to the mean effect. Fortunately, the errors of parameter estimation provided by ANCILLES worked in favor of less biased measurement. The accuracy of the Owen adaptive fixed-length test with this correction was somewhat poorer with parameters estimated by ANCILLES than with known parameters. This drop in accuracy did not appear to be severe enough to discount the Owen procedure for adaptive testing. The Bayes modal adaptive procedure as implemented in this study needs further work to equal or surpass the Owen algorithm, even with more accurately estimated parameters.

REFERENCES

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1979. (NTIS No. AD A067752).

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A047495).

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1977. (NTIS No. AD A044828)

Birnbaum, A. 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, 1968.

Croll, P., & Urry, V. W. ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models, in preparation.

Gorman, S. A comparative evaluation of two Bayesian adaptive ability estimation procedures with a conventional test strategy, in preparation.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary estimation procedure. In C.L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (PS-75-6). U.S. Civil Service Commission, Personnel Research and Development Center, Washington, DC: U. S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Kolakowski, D., & Bock, R. D. A FORTRAN IV program for maximum likelihood item analysis and test scoring: Normal ogive model (Research Memo No. 12). Chicago: University of Chicago, Department of Education, Statistics Laboratory, 1970.

Kolakowski, D., & Bock, R. D. LOGOG: Maximum likelihood item analysis and test scoring: Logistic model for multiple responses (Research Memo No. 13). Chicago: University of Chicago, Department of Education, Statistics Laboratory, 1972.

Learmonth, G. E., & Lewis, P. A. W. Naval Postgraduate School Random Generator Package: LLRANDOM (Research Report NPS55LW73061A) Monterey, CA: Naval Postgraduate School, 1973.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

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

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, 34 (4, Pt. 2, No. 17)

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.

Urry, V. W. A monte carlo investigation of logistic model test models (Doctoral dissertation, Purdue University, 1970). Dissertation Abstracts International, 1971, 31, 6319B. (University Microfilms No. 71-9475)

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Symposium presented at the 83rd annual convention of the American Psychological Association, Chicago, August 1975. Washington, DC: U.S. Civil Service Commission, Personnel research and Development Center, September 1976. (NTIS No. PB 261 694)

Urry, V. W. Tailored testing: A spectacular success for latent trait theory. Springfield, VA: National Technical Information Service, 1977.

Wood, R. L., Wingersky, M. S., & Lord, F. M. Logist: A computer program for estimating examinee ablity and item characteristic curve parameters (Research Memorandum RM 76-6). Princeton, NJ: Educational Testing Service, 1976.