Adaptive EAP Estimation of Ability in a Microcomputer Environment

R. Darrell Bock
University of Chicago

Robert J. Mislevy
National Opinion Research Center

Expected a posteriori (EAP) estimation of ability, based on numerical evaluation of the mean and variance of the posterior distribution, is shown to have unusually good properties for computerized adaptive testing. The calculations are not complex, precede noniteratively by simple summation of log likelihoods as items are added, and require only values of the response function obtainable from precalculated tables at a limited number of quadrature points. Simulation studies are reported showing the near equivalence of the posterior standard deviation and the standard error of measurement. When the adaptive testings terminate at a fixed posterior standard deviation criterion of .90 or better, the regression of the EAP estimator on true ability is virtually linear with slope equal to the reliability, and the measurement error homogeneous, in the range ± 2.5 standard deviations.

With the increasing availability of inexpensive microcomputers, adaptive testing of cognitive abilities is fast becoming a practical reality. Many, perhaps most, applications of mental testing will soon benefit from the flexibility and efficiency of computerized adaptive testing. The requisite statistical theory, including realistic item response models (Samejima, 1981) and rigorous methods of item parameter estimation (Bock & Aitkin, 1981; Reiser, 1982; Thissen, 1982), is now available. Production computer programs based on these principles are nearing completion (Mislevy & Bock, 1982; Thissen, in press).

Adaptive testing depends fundamentally upon item-invariant methods of estimating a person's ability. For items to be chosen sequentially to maximize information gain during an adaptive testing session, it must be possible to obtain comparable estimates of ability from responses to arbitrary sets of items. Several item response theory (IRT) estimation techniques with this property have been proposed.

Birnbaum (1958) was the first to show that maximum likelihood estimates of a person's ability can be obtained from an arbitrary set of items for which continuous response functions with respect to a common dimension can be specified (Birnbaum, 1968). He presented the likelihood equations for the two- and three-parameter logistic models for response probabilities as a function of latent ability and known item parameters. He also introduced the concept of the item and test information functions that make it apparent how to choose items to maximize the precision of estimation at any given level of ability.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 4, Fall 1982, pp. 431–444

0416-6216/82/040431-14$1.70

Later, Samejima (1969) extended these results to more general IRT models and proposed the Bayes modal estimator (maximizing the posterior density of ability, given the examinee's item responses) as an alternative to the maximum likelihood estimator. She did not consider the possibility of using the mean of the posterior distribution (i.e., the Bayes estimator) as the estimate of ability, because at the time it was not clear how the necessary expected values could be obtained.

Recently, however, Bock and Aitken (1981) have called attention to quadrature formulas that enable efficient numerical calculation of the mean and variance of the posterior distribution. They have proposed the practical use of these formulas for test scoring, introducing the terms MAP (maximum a posteriori) and EAP (estimated a posteriori) estimators to distinguish, respectively, the Bayes modal and the Bayes estimators. Unlike the Bayes procedure proposed by Owen (1969), their method does not update the ability estimate after each item response on the patently false assumption that the posterior distribution is normal. Their method is also invariant with respect to the order in which the items enter the calculations, a property not shared by Owen's procedure.

The purpose of the present article is to point out the many advantages of the EAP estimator in computerized adaptive testing where efficiency of computation is at a premium. The EAP estimate computed by quadrature requires significantly fewer operations than the MAP or the maximum likelihood (ML) estimate. The log likelihoods employed in these calculations accumulate as simple sums, as successive items are presented. Moreover, the response probabilities at the assigned quadrature points can be evaluated beforehand and can be stored with the respective item in the item pool. Calculation of the EAP estimates, and the variance of the posterior distribution that serves as its error variance, are thus simple sums of products of fixed quantities. The only additional operation required is an exponentiation to convert the accumulated log likelihood to a likelihood. Because values of the response function are required only at a small number of specified points, a mathematical expression for the function is not required. A graph, fitted in any way to empirical response proportions, could provide these values.

When the response function is complex, as in multiple category scoring (Bock, 1972; Samejima, 1969, 1981), the EAP estimator has the advantage of not requiring calculation of the first derivative with respect to the ability variable. Apart from the need for response probabilities for each distinct category to be precalculated at each quadrature point and stored with the respective item in the item pool, the calculations are the same as for binary scored items.

Among the multiple category models especially promising for adaptive testing, where it is perhaps preferable to require the examinee to respond to every item presented, is Samejima's (1981) modification of Bock's (1972) nominal categories model to allow for effects of guessing. Her model is able to represent the nonmonotonic trace lines that are often seen when the items are extremely difficult and omitting is not permitted (see, for example, the empirical trace lines reported by Bock & Mislevy, 1981). Thissen's (in press) further modification of this model also provides for the possibility that certain of the alternative positions (perhaps depending upon the arrangement of response keys on the console) are more attractive of guesses than others. This model and its derivatives are sufficiently complex that their real-time calculation required in ML and MAP estimation of ability would tax most microcomputers. EAP estimation with this model, on the other hand, would be only negligibly more difficult than with binary models.

Computational Formulas

The computational simplicity of Bock and Aitkin's (1981) method of computing the EAP estimate and the posterior standard deviation is apparent from their formulas: Let

$$x_j = \begin{cases} 1 \text{ if correct} \\ 0 \text{ otherwise} \end{cases} \qquad [1]$$

be the binary score for item *j*, and let

$$P(x_j = 1|\theta) = \Xi_j(\theta) \qquad [2]$$

be the probability that $x = 1$ at the point θ on the ability continuum. Then the likelihood of θ, given the response pattern $[x_1, x_2, ..., x_J]$, is

$$L_J(\theta) = \prod_{j=1}^{J} \left[\Xi_j(\theta)\right]^{x_j} \left[1 - \Xi_j(\theta)\right]^{1-x_j} . \qquad [3]$$

At the J^{th} trial in an adaptive test, the provisional EAP estimate of the ability of person i, $\bar{\theta}_J$, given the item responses, is approximated by

$$\bar{\theta}_J = \sum_{k=1}^{q} X_k L_J(X_k) \cdot W(X_k) \Bigg/ \sum_{k=1}^{q} L_J(X_k) \cdot W(X_k) \qquad [4]$$

and the posterior standard deviation (PSD) is approximated by

$$\mathrm{PSD}(\theta) = \left[\sum_{k=1}^{q} (X_k - \bar{\theta}_J)^2 L_J(X_k) \cdot W(X_k) \Bigg/ \sum_{k=1}^{q} L_J(X_k) \cdot W(X_k) \right]^{1/2} \qquad [5]$$

In Equations 4 and 5, X_k is one of q quadrature points, and $W(X_k)$ is a weight associated with that point. The weights are normed so that

$$\sum_{k=1}^{q} W(X_k) = 1 . \qquad [6]$$

In the context of EAP estimation, the weights are the probabilities at the corresponding points of a discrete prior distribution. In certain cases, for example when a normal prior distribution is assumed, the points and weights can be chosen to improve the accuracy of the numerical approximation of the integral. The Gauss-Hermite points and weights give the exact value of the integral of any function, weighted by the Gaussian error function, that can be expressed as a polynomial of degree *q* (Stroud & Sechrest, 1966). Unfortunately, this class does not include the likelihood functions in adaptive testing, so the Gauss-Hermite quadrature has less advantage in the present application. For the purposes of this article, evenly spaced points in the range ± 3 standard deviations will be employed in the quadratures and the weights will be set equal to the prior discrete probability at these points. Sheppard's correction will be used in calculating the posterior variance to allow for the effects of discretizing the posterior distribution. In application to real populations, perhaps 80 quadrature points between ± 4 standard deviations should be available to insure precision down to PSD = .2. In any given quadrature only about 10 of these points need to be used.

A Simulated Adaptive Testing Session

To illustrate EAP estimation of ability in adaptive testing, 20 responses of an examinee with ability − 0.5 were simulated to items with operating characteristics similar to those of the Word Knowledge Test in the *Armed Services Vocational Aptitude Battery* (ASVAB), Version 8a. Random binary responses of the examinee were generated assuming a three-parameter logistic model. The item parameters for the assumed pool of items are given in Bock and Mislevy (1981). The prior distribution was assumed to be standard normal.

The record of this simulated adaptive testing is shown in Table 1. The prior probabilities (i.e., normalized densities), likelihoods, and posterior probabilities for Items 1, 2, 4, 8, 16, and 20 are also shown in Figure 1. The normalized densities at the 21 equally spaced points have been connected in Figure 1 by smooth curves for illustrative purposes. These curves show clearly the progressive domination of the prior by the data. After 20 items the measurement error (PSD) has been reduced to about .25, the likelihood and posterior density are almost indistinguishable, and the mean and mode of the posterior distribution are almost identical with the ability level at the maximum of the likelihood.

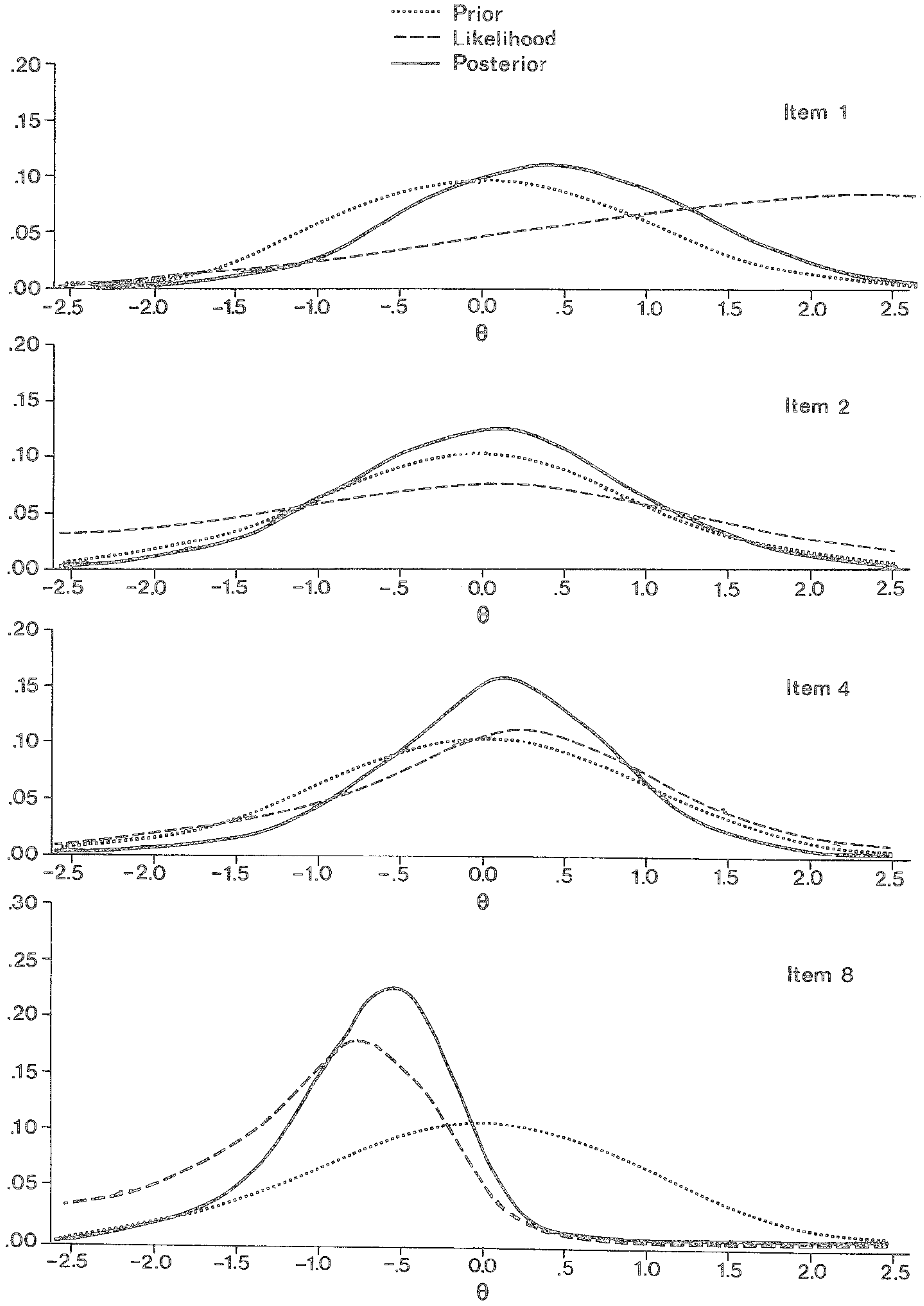
Table 1
Results of a Simulated Adaptive Testing Using
Items From the ASVAB Word Knowledge Test
with Generating Ability = -0.5 and Prior Distribution N(0, 1)

	Selected Item					Posterior	
Attempt	Number	Threshold	Slope	Guessing	Response	Mean	SD
1	24	-.05	.71	.10	1	.36	.89
2	32	.38	.65	.10	0	-.01	.83
3	34	.04	1.11	.10	1	.37	.74
4	29	.39	.69	.10	0	.10	.70
5	35	.14	2.10	.10	0	-.34	.56
6	28	-.42	1.00	.10	0	-.58	.57
7	19	-.54	.93	.10	1	-.40	.50
8	22	-.42	1.68	.10	0	-.66	.48
9	9	-.64	.92	.10	0	-.82	.48
10	14	-.84	1.23	.10	1	-.66	.42
11	17	-.69	1.96	.10	1	-.49	.36
12	25	-.19	.88	.10	0	-.56	.36
13	27	-.16	.82	.10	0	-.62	.36
14	21	-.90	1.03	.10	1	-.55	.33
15	18	-.90	1.23	.10	1	-.49	.31
16	26	-.26	1.68	.10	0	-.58	.30
17	16	-1.01	.95	.10	1	-.54	.29
18	11	-.91	1.34	.10	1	-.49	.27
19	23	-.13	1.63	.10	0	-.54	.26
20	15	-.86	1.68	.10	1	-.49	.25

Reliability of the EAP Estimate

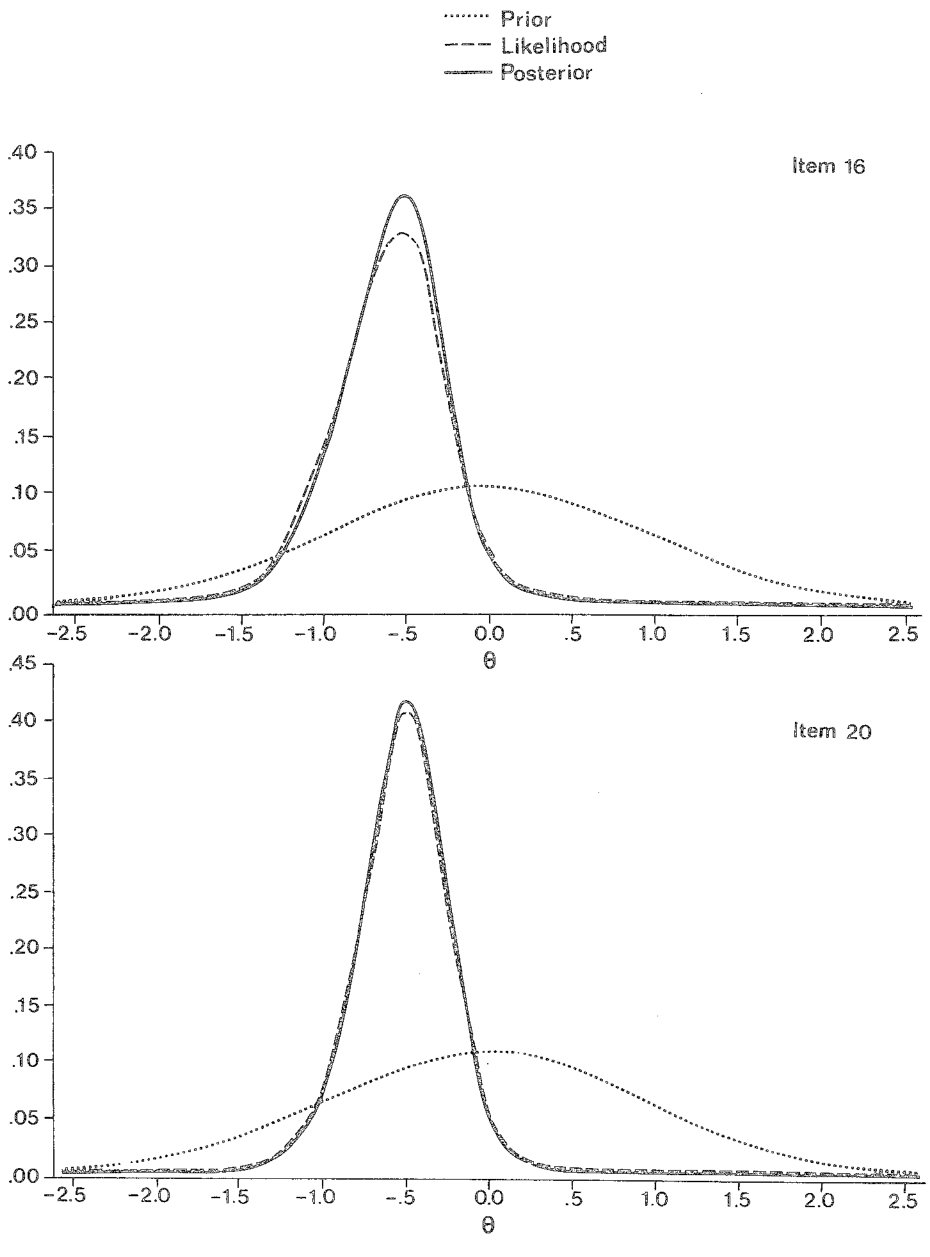
In EAP estimation a measure of the precision of the estimated score is given by the PSD. As the number of items increase, the posterior distribution approaches normality, and its mean and variance

Figure 1
Prior Probabilities, Likelihoods, and Posterior Probabilities for
Selected Items in a Simulated Adaptive Test
of ASVAB Word Knowledge



Continued on the Next Page

Figure 1, Continued
Prior Probabilities, Likelihoods, and Posterior Probabilities for
Selected Items in a Simulated Adaptive Test
of ASVAB Word Knowledge



characterize almost completely the range of true values of ability that is consistent with the observed responses. Thus, the PSD of the posterior plays the same role as the asymptotic standard error (SE) of the ML estimator. The near identity of the likelihood function and posterior destribution after 20 items, shown in Figure 1, illustrates the virtual interchangeability of posterior standard deviations and standard errors in adaptive testing when the number of items reaches 20 and above.

It is assumed in these calculations that ability is distributed $N(0, 1)$ in the population. In this case, the intraclass correlation

$$\rho = 1 - [\mathrm{PSD}(\theta)]^2 \qquad [7]$$

is the reliability coefficient for the EAP estimate. In particular, the reliability of the estimated ability after the 20th item in the adaptive test recorded in Table 1 is $1 - (0.25)^2 = .9375$.

If the adaptive item presentations are continued until a specified PSD is attained, the reliability of the ability estimates is the same for all persons tested for which a sufficient number of items has been administered. Uniform measurement error is an attractive property from many points of view:

1. It satisfies the assumption in classical true score theory of a constant measurement error variance independent of the level of ability.
2. If the estimated abilities are treated as data, it avoids the difficulties for standard statistical procedures that nonhomogeneous error variances present (see Bock & Thrash, 1977, for an example of the complexities of adapting a standard procedure to ML estimates of ability).
3. If the estimates are used for selection and classification of personnel at numerous levels of ability, it insures that the errors of misclassification will be uniform for all decisions.

This type of equitability of selection and classification procedures cannot be achieved by conventional fixed-length tests.

Results of a Large Number of Simulated Adaptive Testing Sessions

In practice, perfect equity will not strictly obtain in adaptive testing if the session has to be terminated before the variance criterion is reached, either because suitable items in the pool are exhausted or because the available testing time is exceeded. To have some idea of the practicality of testing to a specified level of reliability, 500 adaptive testing sessions for both the two- and three-parameter logistic models were simulated at each of three variance limits. The slope parameter of the three-parameter logistic model was assumed equal to 1 and the random guessing parameter equal to .2. The threshold parameter (difficulty) was set equal to the value that is maximally informative at the provisional estimate of ability after each response of the examinee. Birnbaum's (1968) formula (p. 463) was used for this purpose. The starting value for the ability estimate was 0 in all cases. The assumed parameter values are in a realistic range for typical educational and vocational tests.

Table 2 shows the average number of items required to reach the error criterion PSD less than or equal to .4, .3, and .2 under the two models. It is apparent that with items of the assumed discriminating power, a PSD of .3, corresponding to a reliability of about .9, can be obtained with a reasonable number of items. On the other hand, a PSD of .2, corresponding to a reliability of about .95, requires rather large numbers of items, especially when effects of guessing add noise to the data.

Even with a PSD criterion of .3, however, unfavorable cases will occur in which the number of items required to reach this value may be excessive. Table 3 shows the distribution of numbers of items required for this PSD when the guessing model is assumed. About half the cases in 500 required 24 or less items, 95% reached criterion in 30 or less items, but 100% success was attained only by test-

Table 2
Average Number of Items Required to Test Adaptively to Various PSD in Simulated ASVAB Word Knowledge Test Using Two- and Three-Parameter Logistic Models

		Average Number of Items Required	
PSD	Reliability	2-Parameter	3-Parameter
.4	.84	9	15
.3	.91	17	25
.2	.96	37	55

ing with up to 38 items. Although this would not be an excessive number in applications where the examinees are free to proceed at their own pace, it might be considered excessive if time on the testing terminals had to be tightly scheduled. In the latter situation, a slightly lower reliability would have to be accepted in only a small proportion of cases. The results for PSD = .2 in Table 3, on the other hand, suggest that the corresponding reliability of about .95 could be reached in most cases only with a large item pool and generous testing times.

Table 3
Number of Items Required to Test Adaptively to Various PSD in Simulated ASVAB Word Knowledge Tests (500 Tests Each) Using Three-Parameter Logistic Model

Number of Items Required	Number of Testings		
	PSD=.4	PSD=.3	PSD=.2
6-10	2	-	-
11-15	384	2	-
16-20	111	1	-
21-25	3	321	2
26-30	-	160	-
31-35	-	15	1
36-40	-	-	-
41-45	-	-	-
46-50	-	-	2
51-55	-	-	325
56-60	-	-	140
61-65	-	-	20
> 65	-	-	10

Bias of the EAP Estimate

As is well known, the EAP estimator has minimum mean square error over the population of ability and, in terms of average accuracy, cannot be improved upon. At any particular level of ability, however, the expected value of the estimator is not equal to the true ability—that is, the EAP estimator is not unbiased in finite samples of items except at the population mean. Only as the number of items increases without limit does the estimator equal the true ability—that is, the estimator is consistent in the number of items.

In addition, EAP estimates of extreme abilities may be further biased if the initial ability for adaptive testing is set near the middle of the prior distribution. The intermediate estimates then tend to approach the true value from the inside and stop somewhat short of that value when the error criterion is reached. Whether or not these biases are large enough to outweigh the advantages of EAP estimation is an empirical question.

To determine the extent of bias at various error criteria and ability levels, a further set of simulations were carried out. One hundred adaptive testings were simulated at ability levels from − 3 to + 3 in steps of .2 and for error criteria PSD = .4, .3, and .2. The means and standard deviations for the 100 estimated abilities in each array are shown in Table 4. Note that the array standard deviations are very close to the corresponding PSD. This supports the practice of referring to the PSD as if it were a standard error (SE) with its attendent frequency interpretation.

Figure 2
True and Estimated Ability in 500 Simulated
Adaptive Testings to PSD = .4

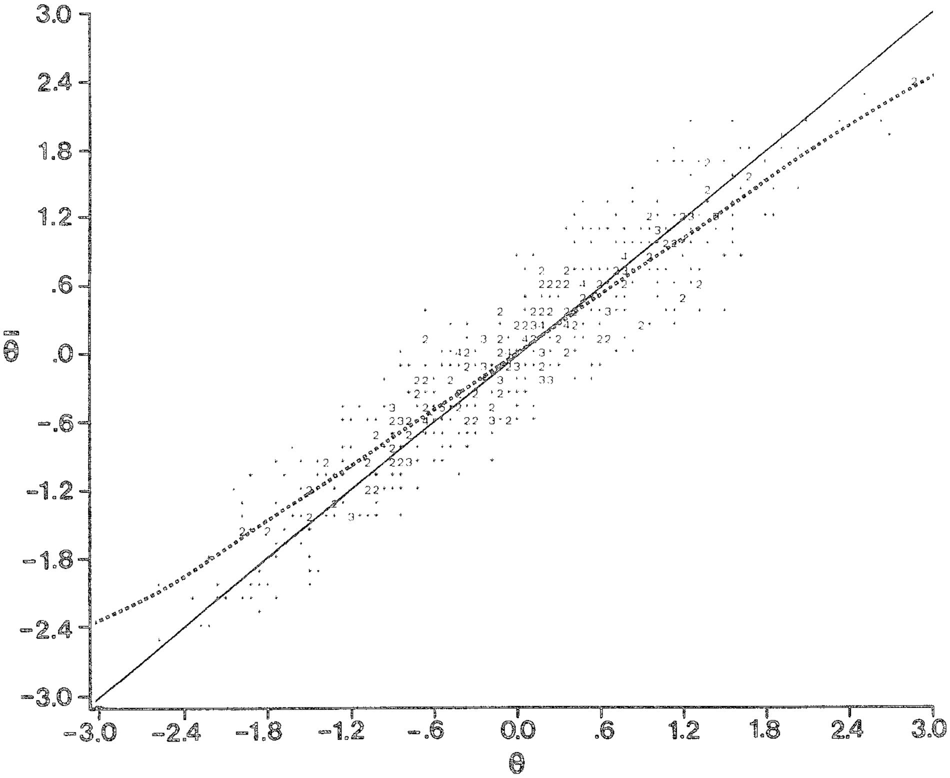


Table 4
Results of 100 Simulated Adaptive Testings at Selected Ability Levels for Three Levels of PSD Using Three-Parameter Logistic Model

Generating	Mean and SD of EAP Estimates					
	PSD=.4		PSD=.3		PSD=.2	
Ability	Mean	SD	Mean	SD	Mean	SD
-3.00	-2.34	.44	-2.65	.30	-2.84	.22
-2.80	-2.28	.48	-2.57	.33	-2.70	.21
-2.60	-2.19	.45	-2.37	.34	-2.52	.21
-2.40	-1.98	.41	-2.15	.32	-2.25	.22
-2.20	-1.87	.39	-2.02	.31	-2.11	.21
-2.00	-1.62	.44	-1.77	.31	-1.91	.21
-1.80	-1.49	.35	-1.64	.25	-1.75	.20
-1.60	-1.36	.40	-1.46	.35	-1.53	.20
-1.40	-1.18	.42	-1.28	.26	-1.37	.16
-1.20	-1.05	.36	-1.14	.34	-1.17	.20
-1.00	-.87	.44	-.93	.29	-.96	.19
-.80	-.64	.31	-.70	.26	-.77	.19
-.60	-.54	.36	-.56	.29	-.57	.19
-.40	-.35	.33	-.37	.30	-.39	.21
-.20	-.14	.38	-.15	.28	-.22	.19
0.00	-.01	.33	.02	.27	.04	.19
.20	.12	.34	.13	.28	.16	.20
.40	.33	.38	.35	.31	.37	.22
.60	.55	.34	.57	.27	.60	.21
.80	.67	.35	.73	.29	.80	.19
1.00	.83	.35	.87	.25	.93	.17
1.20	1.02	.33	1.08	.29	1.16	.20
1.40	1.21	.38	1.27	.31	1.34	.21
1.60	1.44	.31	1.52	.28	1.58	.20
1.80	1.54	.35	1.64	.30	1.74	.18
2.00	1.69	.37	1.83	.29	1.92	.20
2.20	1.89	.35	1.99	.27	2.11	.19
2.40	2.04	.36	2.16	.25	2.29	.18
2.60	2.24	.36	2.38	.27	2.49	.19
2.80	2.29	.40	2.52	.30	2.67	.21
3.00	2.49	.38	2.71	.30	2.86	.20

For the three criterion levels the simulations for the guessing model yielded the estimates plotted in Figures 2, 3, and 4 versus the generating ability. The solid line drawn through the points represents equality of the estimate and the true value, and the dashed line is a smoothed curve connecting the vertical array means. It is apparent in these figures that at reliability level .9 and above, the bias is negligible for 95% of the population, amounting to at most about .1 standard deviations. Only for more extreme θ levels does the bias become appreciable. In Figure 3 the line connecting the array means deviates only slightly from the straight line regression with slope .9 that would obtain if the joint distribution of true and estimated abilities were unit bivariate normal with correlation .9. The plots for PSD = .3 and .2 illustrate how closely the adaptive estimates conform to classical true-score theory. They suggest that this method of estimating ability justifies the many results of classical test theory (Gulliksen, 1950) better than the test scores for which they were intended.

Most of the bias in these simulated results could be eliminated by dividing the deviates from the population mean by the corresponding reliability coefficient. The effect, however, would be to increase the mean square error over the population as a whole. In the practical use of EAP estimates for personnel decisions, such corrections are not required provided two conditions are met: (1) the abilities of persons competing for the same position must be estimated with respect to the same prior distribution and (2) the abilities of such persons must be estimated with the same precision (reliability).

Figure 3
True and Estimated Ability in 500 Simulated Adaptive Testings to PSD = .3

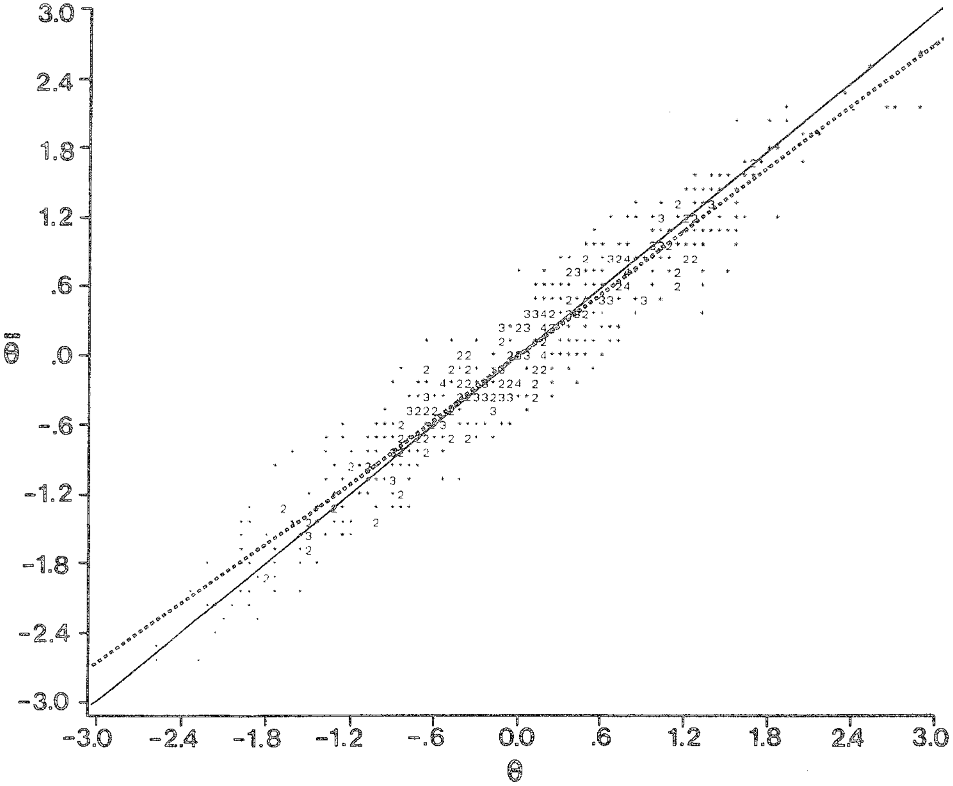
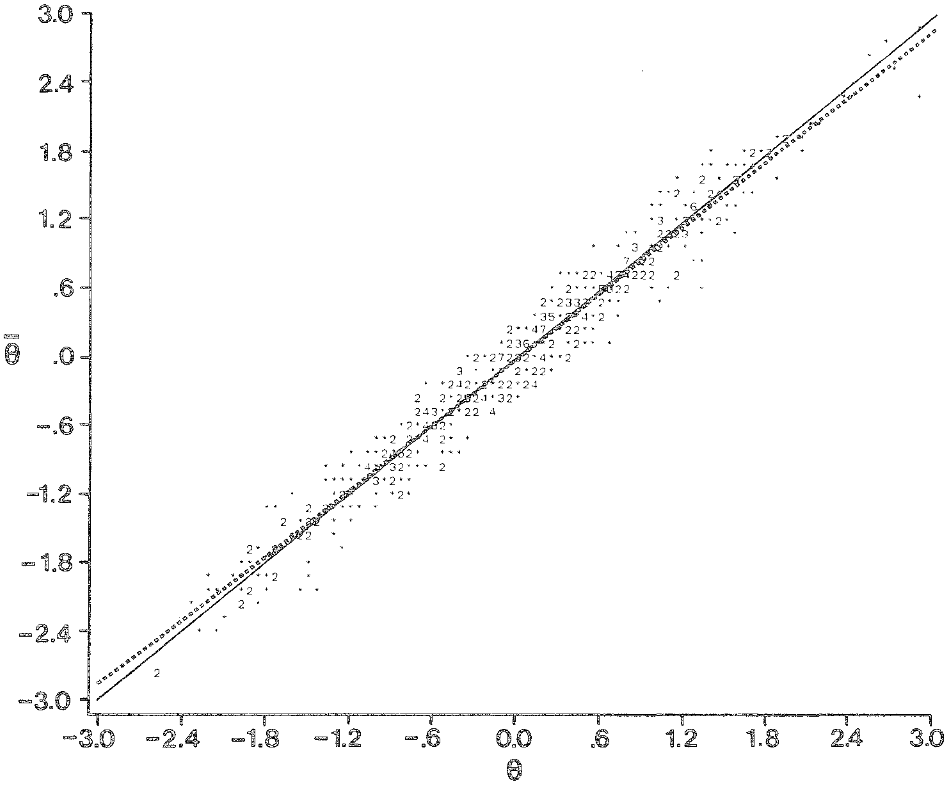


Figure 4
True and Estimated Ability in 500 Simulated
Adaptive Testings to PSD = .2



Assigning the person with the greater EAP estimated scores is then an equitable decision in the sense that any biases that might be present in the estimates are equal for both examinees. In the infrequent cases where the adaptive testing had to be terminated without reaching the fixed error criterion, a correction for the differing reliabilities of the estimates could be made. In true-score theory, this would be equivalent to comparing persons with respect to regressed scores rather than raw scores (see Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Another notable feature of the plots shown in Figures 2, 3, and 4 is the absence of outliers at the extremes of the ability range. Both EAP and MAP estimation yield no values that could not have plausibly arisen from the prior distribution, given the data. Even all-correct and all-incorrect answer patterns give finite and plausible estimates, provided of course that a proper (i.e., nonuniform) prior has been specified. These estimators are much better behaved in this respect than the ML estimator, which does not exist for the all-correct or all-incorrect patterns, and can produce extreme outliers when the response patterns are unfavorable.

This is an especially important consideration in adaptive testing, where a plausible estimate is desirable after every response, including the first. The property is shared by the MAP estimator, but the EAP estimate, based as it is on the mean of the posterior distribution rather than the mode, is more stable than MAP in early phases of an adaptive test.

Summary and Conclusions

The method of calculating expected a posteriori (EAP) or ‘‘Bayes’’ estimation of ability by quadrature with respect to a finite discrete prior distribution, proposed by Bock and Aitkin (1981), is shown to have many advantages over maximum likelihood (ML) and maximum a posteriori (MAP, or ‘‘Bayes modal’’) estimation in computerized adaptive testing.

1. EAP estimates are easy to compute because they are noniterative and require values of the item response functions only at preassigned points on the ability continuum. These values can be calculated externally and stored with the respective items in the pools available to the computer during adaptive testing. The log likelihoods at these points, given the data, accumulate as simple sums as each item is responded to, and the EAP estimate and its variance is computed simply as a mean and variance over the points weighted by the exponential of the respective log likelihood and multiplied by the quadrature weight.
2. Unlike MAP, the EAP estimator does not require the calculation of derivatives of the response function. It can employ a discrete prior distribution, including finite representations that are easy to estimate empirically, thus freeing the testing procedures from arbitrary assumptions about the prior distribution (see Mislevy, 1982).
3. EAP estimates always exist, even for the all-correct or all-incorrect answer patterns and other unfavorable patterns under the three-parameter model for which the ML estimator does not exist. The EAP estimator is therefore stable at all adaptive test lengths, including the first few items administered.
4. No other estimator than EAP has smaller mean square error over the population for which the distribution of ability is specified by the prior.
5. If the adaptive item presentation is continued until a reasonable stringent error criterion is satisfied (e.g., corresponding to a reliability of .9), the bias of the EAP estimate is minor for most of the population (within ± 2 standard deviations). Simulations reported in this paper suggest that this level of reliability can be attained with items typical of cognitive and vocational tests in an average of 25 items and in 95% of cases with 30 items.
6. The bias of the EAP estimator is approximately the same as that of regressed estimates for the same reliabilities and at the same ability levels.
7. Personnel decisions based on EAP estimates are equitable provided persons competing for the same positions are estimated with respect to the same prior distribution and to the same error criterion.
8. Finally, the EAP estimation procedure has the interesting property of employing calculations that are part of the E step of Bock and Aitken’s (1981) EM solution of the likelihood equations for marginal ML estimation of item parameters. The E step consists of accumulating over the sample of persons the likelihoods (1) at each quadrature point and (2) for each item when the person’s response to that item is correct. This same information is available when the adaptive testing of a person in the sample has satisfied the error criterion. If at this point an additional uncalibrated item is presented to the examinee, the likelihoods can be saved by the system and added to the likelihood function resulting from other examinees. When sufficient data of this type has accumulated from persons in the sample, the M step can be carried out to estimate the parameters of the new item with respect to the dimension defined by the existing item pool. If the discrimination parameter of the new item is sufficiently high and the threshold (difficulty) is in a useful range, the item can be added to the pool. In this way, the pool can be updated during operational use of the adaptive testing system and without resort to external calibration studies.

References

Birnbaum, A. *On the estimation of mental ability* (Series Report No. 15, Project No. 7755-23). Randolph Air Force Base TX: USAF School of Aviation Medicine, 1958.

Birnbaum, A. Some latent ability 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.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29-51.

Bock, R. D., & Aitken, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 1981, *46*, 443-459.

Bock, R. D., & Mislevy, R. J., *Data quality analysis of the Armed Services Vocational Aptitude Battery.* Chicago: National Opinion Research Center, 1981.

Bock, R. D., & Thrash, W. A. Characterizing a latent trait distribution. In P. R. Krishnaiah (Ed.), *Applications of statistics.* Amsterdam: North-Holland, 1977.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley & Sons, 1972.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Mislevy, R. J., *Characterizing latent distributions.* In preparation, 1982.

Mislevy, R. J., & Bock, R. D. *BILOG: Item analysis and test scoring for binary logistic models.* Chicago: International Educational Services, 1982.

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

Reiser, M. *Constrained two and three parameter estimation for the normal ogive model.* Submitted for publication, 1982.

Samejima, F. Estimation of latent ability using a pattern of graded scores. *Psychometric Monograph*, No. 17, 1969.

Samejima, F. *Efficient methods of estimating operating characteristics of item response categories and challenge to a new model for the multiple choice item* (Final Report NR 150-402). Office of Naval Research, 1981.

Stroud, A. H., & Sechrest, D. *Gaussian quadrature formulas.* Englewood Cliffs NJ: Prentice-Hall, 1966.

Thissen, D. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 1982, 175-186.

Thissen, D. *MULTILOG: Item analysis and scoring with multiple category response models.* Chicago: International Educational Services, in press.

Author's Address

Send requests for reprints or further information to R. Darrell Bock, Department of Behavioral Science, University of Chicago, 5848 S. University Ave., Chicago IL 60637 U.S.A.