Essentially Unbiased EAP Estimates in Computerized Adaptive Testing

Tianyou Wang

ACT, Inc.

Address correspondence to Tianyou Wang, ACT, P.O. Box 168, Iowa City, IA 52243,
e-mail: wang@act.org

Essentially Unbiased EAP Estimates in Computerized Adaptive Testing

Abstract

In computerized adaptive testing (CAT), the scoring procedure is usually based on IRT-based ability (θ) estimates instead of number-correct scores because different examinees typically receive different sets of items. It is well-known that the maximum likelihood estimation (MLE) produces relatively unbiased estimates with relatively high standard error (SE) in CAT. The Bayesian estimation methods, on the other hand, produce estimates with relatively small SE but with large bias if a standard normal prior is imposed. The purpose of this paper was to propose a new expected a posteriori (EAP) estimation method with a flatter prior distribution than the standard normal distribution to reduce the bias of the Bayesian methods. The simulation results of the paper demonstrated that the EAP with a beta prior distribution can produce estimates with similar or even smaller bias than the MLE and yet does not sacrifice much of the smaller SE and root mean square error (RMSE) of the standard EAP estimation with a normal prior, and that the presence of practical constraints such as content balancing and item exposure rate control does not affect the relative unbiasedness of the new EAP method.

Key Words: Computerized adaptive testing, Bayesian estimation, expected a posteriori, prior distribution, bias.

Essentially Unbiased EAP Estimates in Computerized Adaptive Testing

In computerized adaptive testing (CAT), it is common that different examinees receive different sets of items from a given item pool. Because those sets of items are of different difficulty levels, it is inconvenient to derive the reported scores based on the number-correct raw scores as is often done in paper-pencil conventional testing. Therefore, IRT-based ability (θ) estimates are often used as a basis in deriving the reported scores. So far, four ability estimation methods primarily have been used in CAT: (1) maximum likelihood estimation (MLE) (Birnbaum, 1968), (2) Owen's Bayesian estimation (OWEN) (Owen, 1969, 1975), (3) expected a posteriori estimation (EAP) (Bock & Aitken, 1981; Bock & Mislevy, 1982), and (4) maximum a posteriori estimation (MAP) (Samejima, 1969). A few studies (Bock & Mislevy, 1982; Weiss & McBride, 1984; de la Torre, 1991; Wang, 1995) have been done to examine and compare these ability estimation methods under CAT settings. The general conclusions are that MLE is relatively unbiased with a well-designed item pool but has relatively large standard error (SE), and that the Bayesian methods are relatively biased toward the prior mean, and that among the Bayesian methods, EAP has relatively small bias, and SE. Bias in this context is defined as the mean θ estimates for an examinee taking the same CAT many times without practice effect minus his/her true θ. EAP has the advantage of being computationally simpler than MLE and MAP. Wang (1995) also found that if an item pool lacks items of extreme difficulty levels, which is usually the case with real-world item pools, MLE could also be biased, but in the opposite direction of the Bayesian methods.

In many standardized testing programs (e.g., GRE, see Eignor & Schaeffer, 1995), Bayesian methods are not used despite their small standard error only because they are seriously biased. Bias can be problematic when the estimates are used to make inferences in relation to some absolute criterion. For instance, in computerized mastery testing, the estimates may be used to compare with certain cut-scores and make decisions about examinees' pass/fail status. Bias in the estimation can cause serious false decisions. For

some testing programs, the CAT form will co-exist with its paper-pencil conventional form for a period of time and the score scale will remain the same as for the conventional form. In these situations, there is a need to transform the θ estimates into the equivalent number-correct score on some base conventional form (e.g., Eignor & Schaeffer, 1995). Any bias in the θ estimates will necessarily affect the transformed reported score in a negative way. To solve this bias problem, some CAT developers resorted to traditional equating methods to eliminate the effect of the bias. For example, Segall (1995) and Segall & Carter (1995) used a random groups design to eliminate the inequivalency of the θ based CAT scores and conventional form test scores. The equating process is usually expensive and may introduce additional errors in the process of data collection and analysis.

Conceptually, the Bayesian methods are intrinsically biased because of the incorporation of the prior information into the estimation process. Like the regression methods in predication problems which regress the predicted values toward the mean, the Bayesian methods also regress estimates toward their prior mean. The Bayesian methods use both the data and the prior for estimation whereas MLE uses only the data. The Bayesian methods can be thought, in some loose sense, as a combination of MLE and the prior distribution which is usually the standard normal distribution. (For convenience, Bayesian methods with a standard normal prior will be called standard Bayesian methods in the remainder of this paper.) The large bias of the standard Bayesian methods in CAT is caused by the steep shape of the standard normal prior. But because MLE also has a relatively small bias in the opposite direction of bias of the Bayesian methods, it was hypothesized that if a flatter prior distribution is specified, the Bayesian estimates can also be relatively unbiased. The purposes of this paper are to study the effect of different specifications of the prior distribution on the bias of the EAP estimates in CAT, and to search for an optimal prior for a given item pool so that the EAP estimates would be basically unbiased in a relatively wide range of the θ scale. The relationship between the characteristics of item pool and the shape of the optimal prior will be investigated. Another purpose of the study is to examine

the possible effects of implementing practical constraints such as content balancing and item exposure rate control on the bias of the new EAP method.

EAP was chosen among the Bayesian methods because of its relatively small error and computational simplicity over other Bayesian methods even though similar idea can be applied to MAP; that is, a flatter prior distribution can be applied to MAP to reduce its bias. Because OWEN was specifically designed to have a standard normal prior, however, this idea does not apply to the OWEN method.

The MLE and EAP Estimation Methods

In order to facilitate the presentation of the new EAP method, the technical aspects of the IRT ability estimation methods, in particular, of MLE and the standard EAP methods are described below.

Maximum likelihood estimation: MLE is a widely used for parameter estimation in many statistical applications. In the context of item response theory (IRT) ability estimation, given a response vector **u** to a set of items with known parameters, the likelihood function is

$$L(\mathbf{u}|\theta) = \prod_{i=1}^{n} P_i(u_i|\theta) \qquad (1)$$

The MLE of θ is $\hat{\theta}$, the value that maximizes this likelihood function (or equivalently, the log of it), as expressed by

$$lnL(\mathbf{u}|\theta) = \sum_i \ln P_i(u_i|\theta) = \sum_i \ln P_i^{u_i} Q_i^{1-u_i} \qquad (2)$$

This can be done by setting the derivative of the likelihood function to zero and solving the resulting equation.

$$\frac{\partial lnL(\mathbf{u}|\theta)}{\partial\theta} = \sum_i \frac{P_i'(u_i|\theta)}{P_i(u_i|\theta)} = \sum_i \frac{(u_i - P_i)P_i'}{P_i Q_i} = 0 \qquad (3)$$

Iterative numerical methods such as the Newton-Raphson method can be used to solve the likelihood function. Asymptotically, the variance of the MLE estimates can be approximated by the inverse of the test information function.

$$Var\left(\hat{\theta}|\theta\right) \approx \frac{1}{I\left\{\theta,\hat{\theta}\right\}} = \frac{1}{\sum_i \frac{P_i'^2}{P_i Q_i}} \tag{4}$$

In the context of CAT, the approximation may not be sufficiently accurate because the test length of a CAT test is supposed to be relatively short. Warm (1989) and Wang (1995) found with simulation that the information-based error variance estimates underestimate the actual error variances.

Lord (1983) derived the bias function for MLE $\hat{\theta}$ as:

$$Bias\left(MLE(\theta)\right) \approx \frac{D}{I^2}\sum_{i=1}^{n} a_i I_i\left(\phi_i - \frac{1}{2}\right), \tag{5}$$

where D=1.7, $I_i = P_i'^2 / P_i Q_i$, $\phi_i = \frac{P_i - c_i}{1 - c_i}$. The bias function suggests that when a set of items are all targeted at an examinee's true ability level, the bias will be close to zero because the term in the parentheses will be close to zero. If the ability level is higher than the average item difficulty level, the bias will be positive; likewise, if the ability level is lower than the average item difficulty level, the bias will be negative.

The expected a posteriori estimation: In the context of ability estimation in IRT, we have

$$p(\theta|\mathbf{u}) = \frac{L(\mathbf{u}|\theta)g(\theta)}{P(\mathbf{u})} = \frac{L(\mathbf{u}|\theta)g(\theta)}{\int L(\mathbf{u}|\theta)g(\theta)d\theta}, \tag{6}$$

where $g(\theta)$ is the prior information of the examinee's ability. The EAP method is to find the mean of the posterior distribution. The mean and the variance given a posterior distribution $p(\theta|\mathbf{u})$ are expressed as follows:

$$E(\theta|\mathbf{u}) = \int_{-\infty}^{\infty} \theta p(\theta|\mathbf{u})d\theta, \tag{7}$$

$$Var(\theta|\mathbf{u}) = \int_{-\infty}^{\infty} \theta^2 p(\theta|\mathbf{u})d\theta - \left(E(\theta|\mathbf{u})\right)^2 \tag{8}$$

Using numerical quadrature (see Stroud & Sechrest, 1966), these integrations can be approximated to any practical degree of accuracy by:

$$\hat{\theta} \equiv E(\theta|u) = \frac{\sum_{k=1}^{q} X_k L(X_k) W(X_k)}{\sum_{k=1}^{q} L(X_k) W(X_k)} , \quad (9)$$

$$\hat{\sigma}^2(\hat{\theta}) = Var(\theta|u) = \frac{\sum_{k=1}^{q} (X_k - \hat{\theta})^2 L(X_k) W(X_k)}{\sum_{k=1}^{q} L(X_k) W(X_k)} , \quad (10)$$

where X_k is one of q quadrature points, $W(X_k)$ is a weight associated with the quadrature point , and $L(X_k)$ is the likelihood function conditioned at that quadrature point. Using this procedure, it can be seen that the EAP estimates become summations and do not require iterative processing. Unlike the OWEN method, the EAP method evaluates the actual posterior distribution directly. So at least logically, the EAP method is superior to the OWEN method. Bock & Mislevy (1982) pointed out that among all possible estimators, EAP has the smallest mean square error over (RMSE) the population for which the distribution of the ability is specified by the prior. The bias of the Bayesian methods all point toward the middle point of the θ scale if a standard normal prior is used. The shape of the prior distribution affects the magnitude of the bias for the Bayesian estimates. In large-scale standardized testing, it is often not realistic to use any actual prior information about the individual examinee to form the prior. Therefore it is common practice to use the standard normal distribution as the prior for every examinee. The bias of the Bayesian estimates represents a regression effect toward the group mean which is undesirable in most standardized testing settings.

The New EAP Estimation Method

The primary feature of the new EAP method proposed here is that a flatter shaped prior distribution is used instead of the standard normal prior. The goal is to make the new EAP method as good as MLE in terms of bias and still have RMSE similar to the standard EAP. With this new EAP method, the prior distributions no longer aims at reflecting any prior information about the examinees' ability but only at serving as a tool to achieve

technical quality such as less bias. For this reason, they can be referred to as uninformative priors.

In choosing such flatter priors, there may be many different options. One such option may be the normal distribution with variance greater than one. But because the magnitude of bias of EAP or other Bayesian methods were found to be generally asymmetric around the middle point of the θ scale (cf. Wang, 1995), the normal distribution is considered not desirable due to its symmetry. The family of beta distributions was considered to be the best option for this situation because of their flexibility in shape. Let this beta distribution be denoted as $g(\theta|\alpha,\beta,l,u)$, where $\alpha,\beta,l,$ and u are four parameters that characterize the distribution, with the first two parameters characterizing the shape and the last two parameters characterizing the lower and upper bounds of the distribution. The probability density function of this distribution can be expressed as (Johnson & Kotz, 1970; Hanson, 1991)

$$g(\theta|\alpha,\beta,l,u)=\frac{(\theta-l)^{\alpha-1}(u-\theta)^{\beta-1}}{B(\alpha,\beta)(u-l)^{\alpha+\beta-1}}\ . \qquad (11)$$

The shape of this distribution is symmetric when α equals β and is asymmetric otherwise. When α is greater than β, it is negatively skewed; otherwise it is positively skewed. The smaller the α and β are, the flatter the shape is. Hanson (1991) presented formulas for computing the mean, variance, skewness and kurtosis of the distribution based on the values of these four parameters. The main task of this study is to search for a way for finding the four parameters so that the resulting EAP estimates will be essentially unbiased along a wide range of the θ scale.

In CAT the bias of the Bayesian estimates is not only affected by shape of the prior distribution, but are also affected by the characteristics of the item pool such as the number of items and the discrimination values within different strata of difficulty levels (Wang, 1995). Therefore a universally applicable prior distribution to produce the least biased estimates for all types of item pools can not be found. The search for such prior distributions

is then specific for a particular item pool. Because many different aspects of the characteristics of an item pool are expected to influence the bias of the EAP estimates (Wang, 1995), it is not expected that parameters for the beta prior can be determined quantitatively in relationship with some indexes of the item pool characteristics. The different aspects of item pool characteristics may include the pool size, the mean discrimination parameter values, the distribution of the difficulty parameters, and the number of items and the mean discrimination values for items within each strata of the difficulty levels, etc. For this reason, a trial-and-error approach with simulations will be used to find the parameter values of the beta prior for a particular pool that yields estimates with the smallest bias. The parameter values thus found, however, will be examined in relationship to the characteristics of the item pool. This process will be repeated across several item pools with different characteristics with the goal of finding general relationships between the parameters of the beta prior distribution with the characteristics of the item pools.

Method and Data

Computer simulation was used in this study. A CAT simulation computer program in C language was developed to simulate and analyze the test data. The simulation was conditioned at each of 17 equal-spaced points on the θ scale from -3.2 to 3.2 in increments of .4. A simulated CAT test of 30 items was replicated over 400 simulees at each of the 17 points. The maximum information item selection algorithm was used to select items from the pool. It was expected that the lower and upper bound of the beta prior, l and u, would not affect the bias as long as they are set with a wide enough range. For this reason, they were fixed at -6 and 6, respectively. A trial-and-error approach was used to determine α and β with many simulations. At first, α and β were both set at 2, which makes the beta distribution very flat. The resulting bias was plotted. Then α and β were both increased in increments of .5. The resulting bias plots were compared to find the parameter values that yielded the smallest bias. If the bias displayed asymmetry, then one of the two parameters was changed to make the bias more symmetric. Once the desired prior distribution was

found (i.e., the smallest bias is achieved), it was compared with the standard EAP estimates and with MLE. Three error indexes were used for comparison: SE, bias, and root mean square error (RMSE). These error indexes were computed conditioned at each of the 17 points on the θ scale with the 400 simulees. The conditional errors were plotted for visual comparison. They were also integrated over a standard normal θ distribution to compute the population overall error indexes. If the EAP with a beta prior produces smaller bias than both MLE and standard EAP and at the same time produces about the same RMSE as standard EAP, then the results were considered satisfactory.

Because it was hypothesized that the condition of the item pool would have a major effect on the shape of the prior and the bias of the new EAP method, the conditions of the item pool vary along two dimensions: the evenness of the item pool quality along the ability scale and the size of the item pool. Two types of item pools were used for simulations: real item pools and generated item pools. Within each type, there are two item pools with different pool sizes: 420 items and 120 items, representing large item pools and small item pools. The real item pools consisted of 420 or 120 ACT Assessment Mathematics items that were calibrated using data collected for equating using a random groups design. The generated item pools consisted 420 or 120 items with generated item parameters. The b parameters for the generated pools were 420 or 120 equally spaced points from -4 to 4. The a parameters were draw on from a normal distribution with mean of 1.2 and s.d. of .2. The c parameters were generated from a truncated normal distribution with a mean of .15 and a s.d. of .05. The real item pools used in this study possess the common characteristics of most real world item pools; i.e., there are more items with medium difficulty levels than items with extreme difficulty levels, and the items with medium difficulty levels have higher discrimination values than items with extreme difficulty levels. In other words, the quality of the real item pools are uneven along the θ scale. With smaller item pool sizes, this problem becomes even more serious. In contrast to the real world item pools, the generated item pools here have even qualities along the entire θ scale. So altogether, four items pools were

used to represent two pool sizes and two categories of item pool characteristics. Table 1 contains the descriptive statistics of the four item pools.

Another important objective of this study was to examine the effect of the presence of practical constraints such as content balancing and item exposure rate control. CAT research in recent years has studied beyond a "pure" CAT and has increasingly focused on such practical constraints (e.g., Stocking & Swanson, 1993; Stocking & Lewis, 1995; Sympson & Hetter, 1985). It is also of interest to investigate whether the imposing of content balancing and item exposure rate control will significantly change the shape of the desirable prior distribution and the change the magnitude of the bias of the new EAP method. Only one item pool, namely, the 420-item ACT Mathematics pool was included to investigate this issue.

The items in the 420-item ACT Mathematics pool belong to six different content categories of the table of specifications. The desirable numbers of items in all of the six content areas for a 30-item test are 7, 4.5, 4.5, 5, 7, and 2, respectively. The Stocking and Swanson's (1993) weighted deviation algorithm was used to achieve satisfactory content balancing. Conditioned at each ability level, a deviation in the mean number of items in each category that is less than one item was considered satisfactory. The Sympson and Hetter's (1985) algorithm was used to control the item exposure rate. The maximum rate was set at 15% percent, i.e., an item can not be exposed to more than 15% percent of the examinees in the population. Because the Sympson and Hetter algorithm is a probabilistic procedure, some items will have exposure rate slightly higher than 15%. The investigation took the following steps. First, the same beta prior for the pure CAT was used for the constrained CAT; Second, if necessary, the beta prior was adjusted to achieve the smallest bias. In order to separate the possible effects of content balancing and exposure rate control, the simulations were repeated with imposing only one of the constraints and with imposing both constraints.

Results

Without Practical Constraints

The simulations were carried out as described above. The optimal α and β parameters of the beta prior were determined using the trial-and-error approach. From the simulation process, it was observed that a difference of .5 in the values of α and β has very little influence on bias of the EAP estimates; therefore it is unnecessary to fine tune the α and β values more than they were in this process. The α and β values thus found are contained in Table 2. It can be seen from Table 2 that item pool size has little effect on the α and β parameter values. The other aspect of the item pool characteristic, namely whether the pool quality is even along the θ scale, however, seems to have a major impact on the values of α and β parameters. The α and β values are much higher with the real item pools than with the generated item pools.

The three error indexes, bias, standard error, and root mean square error (RMSE) are plotted for the three different estimation methods and for the four item pools in Figure 1 through 4. The bias plot in Figure 1 shows that for the large real item pool, MLE has a relatively small bias towards the extremes of the θ scale, whereas the standard EAP has a large bias toward the middle. The new EAP is essentially unbiased in a range of -2.4 to 2.4, with the bias of it starting to increase at the ends of the θ scale. Overall, the new EAP method has smaller bias even than MLE. The SE plot in Figure 1 shows that the new EAP has slightly higher standard error than the standard EAP, but has substantially smaller standard error than MLE. The RMSE plot in Figure 1 shows that the new EAP has slightly higher RMSE than the standard EAP in the middle range of the θ scale but has substantially smaller RMSE than MLE, particularly at the two ends of the θ scale.

Figure 2 shows that for the small real item pool, the errors display similar comparative patterns as for the large real item pool, except that the error indexes were all larger for three methods, particularly for MLE and for the standard EAP. This occurred because the smaller pool size further exaggerated the scarcity of high quality items at the extreme difficulty levels.

The bias plot in Figure 3 shows that with the large generated item pool, the bias of standard EAP is still quite large and is almost linearly related to θ. The new EAP and MLE, however, are consistently unbiased all along the entire θ scale. The SE plot in Figure 3 shows that the standard error for MLE is still larger than that of the standard and new EAP methods, but the discrepancy is smaller than those of the real item pools, particularly at the extreme ends of the scale. The RMSE plot in Figure 3 shows that the new EAP overall still has consistently smaller RMSE than MLE. The RMSE of the standard EAP display a rather flat Vshape due to the high bias at the ends of the θ scale.

Figure 4, which plots the error indexes for the small generated item pool, shows similar comparative pattern to Figure 3. But the magnitude of errors are larger than that displayed in Figure 3. The new EAP still displays smaller bias than the standard EAP but similar SE and RMSE. The standard EAP method is still biased with a magnitude slightly larger than that observed in large generated item pool. The RMSEs for the three methods are quite close to each other except that the standard EAP has large RMSE toward the ends of the θ scale.

A comparison between Figure 1 and 2 with Figure 3 and 4 show that the characteristics of the item pools, particularly the evenness of the quality along the θ scale, may have major impact on the bias and other error indexes of the estimation methods, particularly at the ends of the θ scale. The consistent advantage of the new EAP over the standard EAP in reducing bias and over MLE in reducing SE and RMSE suggest that the usefulness of the new EAP is probably generalizable to almost all real world item pools.

The overall error indexes for the unconstrained CATs are summarized in Table 3a. The overall indexes were computed by numerically integrating the conditional indexes over a normally distributed θ distribution. For the bias index, the absolute values of the conditional bias were used to assess the overall magnitude of the bias. Consistent with the graphical results, Table 3a shows that overall, the new EAP method has similar magnitude of the bias as MLE. More specifically, with the large real item pool and the small generated pool, the overall magnitude of bias of the new EAP is slightly smaller than that of MLE. The

opposite is true for the other two pools. The differences are so small that they are practically negligible. The overall SE and RMSE for the new EAP method are always between those of the standard EAP and MLE.

With Practical Constraints

The implementations of the content balancing and item exposure rate control were successful. The deviations of the mean number of items to the target number in each category were all less then one item. The item exposure rate for all the items was under 16%. The plots of the error indexes for the constrained CAT with the new EAP were contained in Figure 5. The prior distribution here is the same as used in the unconstrained CAT. It can be seen from Figure 5 that the bias of the new EAP method was not affected much by imposing content balancing, or item exposure rate control, or both. Because the adjustment in the α and β parameters of the beta distribution did not improve the bias of the estimates, it was not considered necessary to adjust the beta prior. The SE and RMSE for the new EAP were slightly increased when exposure rate control or both constraints were imposed. But they were not affected much when content balancing was imposed. This result is probably due to the large item pool size and the rather severe exposure rate control imposed in this case. It was also observed that there was a negative bias along the θ scale except for the extreme lower end. This consistent negative bias can be adjusted by adding a small amount to the final ability estimates for every examinee. The bias after this minor adjustment was plotted in Figure 6. SE was not affected by this adjustment and RMSE were slightly reduced for most of the θ scale. For this reason, they were not plotted. Overall, we can conclude that the usefulness of the proposed new EAP method was not affected by the presence of practical constraints such as content balancing and item exposure rate control.

Table 3b summarizes the overall error indexes for the new EAP for different constraint conditions. Consistent with Figure 5, Table 3b shows that exposure rate control increases the overall error indexes more than content balancing. Again, this result may be unique to the conditions in this particular study.

Conclusion and Discussion

Taken as a whole, this study shows that an essentially unbiased EAP estimation with uninformative prior can be found for various item pools in CAT, and that the beta distribution family can be used to serve as such an uninformative prior. Even though an analytic approach for finding the parameters for the beta distribution seems unattainable, the trial-and-error simulation approach used in this study is quite easy to implement. The new EAP basically corrects the severe bias of the standard EAP method without sacrificing much of the low SE and RMSE of the standard EAP. The parameters of the optimal beta prior are closely related to the evenness of the quality of the item pool along the entire θ scale. For item pools with uneven quality, higher values for the shape parameters, α and β, will be required than otherwise. The values of the shape parameters seem not to relate strongly with the size of the item pools.

It should be noted that with the realistic item pools, the relative unbiasedness of the new EAP method is effective only in a range of the θ scale. This effective range may be sufficient for most of the decision purposes in standardized testing. But the presence of this effective range should be kept in mind when this new EAP method is actually applied in CAT testing.

The presence of the practical constraints seems not to affect the bias of the new EAP method much. In this case, adjustment in the shape parameters α and β of the beta prior was not even needed. It is expected that in cases when those adjustments are needed, the amount of adjustment should be fairly small. Other error indexes such as SE and RMSE are only slightly affected by the practical constraints. Such effects are expected to be present for MLE and the standard EAP method as well. In short, under practical constraints, the proposed new EAP method is still a viable solution to the large bias of the standard EAP and the large SE of MLE.

It should be noted that the general idea examined in this paper for the EAP method should also apply to other Bayesian methods such as the MAP method. Whether this idea

would work equally well or even better for MAP than for EAP requires further empirical investigation. Also, the idea of using uninformative prior in Bayesian estimation may apply to other testing situations such as fixed form testing, where IRT estimation techniques are used in scoring.

The ability estimation is an important component of CAT. The bias of the Bayesian methods and the large standard error of MLE are the old problems in constructing practically useful CAT tests. The results of this study seem to present a solution to this important issue which is not difficult to implement in practice.

References

Birnbaum. A. (1968). 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.

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

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Crichton, L. I. (1981). Effect of error in item parameter estimates on adaptive testing. Ann Arbor, Michigan: UMI Dissertation Information Service.

de la Torre (1991). *The development and evaluation of a system for computerized adaptive testing.* Unpublished doctoral dissertation, University of Iowa.

Eignor, D. R. & Schaeffer, G. A. (1995). *Comparability studies for the GRE General CAT and the NCLEX using CAT*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1995.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes. ACT Research Report, 91-5.* ACT, Iowa City, Iowa.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions-2.* New York: Houghton Mifflin.

Lord, F. M. (1983). Unbiased estimators of ability parameters, their variance, and of their parallel-forms reliability. Psychometrika, 48, 233-245

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph, No. 17.*

Segall, D. O. (1995). *Equating the CAT-ASVAB: Experiences and lessons learned*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1995.

Segall, D. O., & Carter, G. (1995). *Equating the CAT-GATB: Issues and approach*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1995.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing. Research Report RR-95-25.* Princeton, NJ.; Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of 27th annual meeting of the military testing association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

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

Wang, T. (1995). *The precision of ability estimation methods in computerized adaptive testing.* Unpublished doctoral dissertation. Iowa City, IA: University of Iowa.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54*, 427-450.

Weiss, D. J. & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273-285.