International Journal of Innovative
Computing, Information and Control
Volume **8**, Number **10(B)**, October **2012**

 ISSN 1349-4198
pp. 7427–7435

A BAYESIAN APPROACH IN MAKING MASTERY DECISIONS: COMPARISON OF TWO LOSS FUNCTIONS

Mingchuan Hsieh

Research Center for Testing and Assessment
National Academy for Educational Research
No. 2, Sanshu Rd., Sansia District, New Taipei City 23703, Taiwan
mhsieh@mail.naer.edu.tw

Received April 2011; revised August 2011

Abstract. *The purpose of this study was to apply a Bayesian procedure in the context of fixed-length mastery tests. The Bayesian procedure was compared with two conventional procedures (conventional-Proportion Correct and conventional-EAP) across different simulation conditions. Two loss functions, including linear and threshold loss function were manipulated. The results show that the Bayesian procedure appeared to effectively control false negative and false positive error rates. The differences in the percentages of correct classifications and phi correlations between true and predicted status for the Bayesian procedures and conventional procedures were quite small. However, there was no consistent advantage for either the linear or threshold loss function.*
Keywords: Bayesian, Mastery decisions, Loss functions

1. **Introduction.** Many test applications involve sorting examinees into two categories, often called masters and non-masters, based on whether or not the examinees have sufficient knowledge and understanding of a particular content domain [5]. Today, mastery tests are applied widely in many fields.

A conventional procedure to determine an examinee's mastery status is to compare the examinee's performance with the cut point. An examinee is declared a master if his/her score on an achievement test is as high as or higher than the pre-specified cut point, or an examinee is declared a non-master if his/her score on the test is lower than the cut point. With the demands for more efficient and accurate decision making, new testing procedures have been developed.

Bayesian approach has been applied widely in many fields [2,6,9,11]. In application to mastery tests, Bayesian method is one of the newly developed procedures and it has drawn some attention for its ability to take the relative costs of erroneous testing outcomes into consideration explicitly. That is, the relative seriousness of false negative errors, false positive errors or administering additional items can be considered in advance in the Bayes procedure by specifying the appropriate values, which are called loss parameters [15,17,18].

In some cases, the test users may desire to minimize a specific kind of classification error. For example, when utilizing board exams to select medical specialists, it is possible that the test user intends to avoid the risk of false positive errors, i.e., selecting examinees whose actual ability levels are below the pass level but declared masters. In this specific condition, the test user can set a higher weight for the loss parameter for false positive errors and a smaller loss weight for false negative errors to decrease the chance of false positive errors in the Bayes procedure.

[3] proposed a Bayesian procedure (named Adaptive Sequential Mastery Testing) which employed the Rasch model to make mastery decisions, which could potentially overcome the drawbacks inherent in mastery decisions based on traditional, non-adaptive tests. Based on [3]'s discussion, the most challenging part in implementing the Bayesian procedure for mastery testing is calculating the expected loss function from the posterior distribution. The posterior distribution involves a prior multiplied by a likelihood function, with the prior typically assumed to follow the standard normal distribution for a Bayesian model in achievement tests. If the likelihood is an exponential function, such as the item response model, e.g., Rasch model, then integrating the posterior distribution over a certain interval to obtain the expected loss function can be challenging.

When mastery testing involves many examinees and items, a larger number of possible response patterns could make the posterior distribution function quite complicated. In the Rasch model, the expected loss function of the posterior distribution can be explicitly estimated since the number correct score is the minimum sufficient statistic for theta. [3] derived a general form to estimate the probability of response patterns by incorporating the sufficient statistics into the exponential function of the Rasch model. The general form they derived could possibly extend to the 2-parameter (2PL) item response theory model; however, their framework has some difficulties in applications of the 3-parameter (3PL) IRT model because of the lack of sufficient statistics.

The purpose of this study is to extend [16]'s Bayesian procedure from the Rasch model to the 3PL IRT model. In addition, the effectiveness of controlling the false positive errors and false negative errors by setting threshold and linear loss function in the Bayesian procedure is investigated. Conventional methods are used to serve as the baseline to evaluate the performance of the Bayesian procedure.

2. **Theoretical Background.**

Before illustrating the mastery procedures used in this study, it is important to review the item response model first.

2.1. **Item response model.**

The item responses used in this study were generated based on the three-parameter logistic model (3PL), which is commonly used in the measurement field. In this model, three item parameters are considered-discrimination, difficulty and guessing. The probability function for the 3PL model can be expressed as follows:

$$P_{ij}(X_{ij}|a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i) * \frac{1}{1 + \exp(-1.7 * a_i * (\theta_j - b_i))}$$

where P_{ij} is the probability of a correct response on item i for a person j; a_i, b_i, c_i are the discrimination, difficulty and pseudo-guessing parameters, respectively for item i; θ_j is an ability parameter for person j.

2.2. **Mastery procedures.**

2.2.1. *Conventional procedure-proportion correct.* In this context of mastery tests, it is common to define the criterion for mastery using the percentage of the items on the test correctly answered. After all items in the conventional test are administered, if the examinee's score is equal to or exceeds the proportion correct cut score, the examinee is declared a master. Otherwise, the non-master decision is declared. In some cases, if IRT is used, conversion of the latent trait score metric to the proportion correct score scale may be necessary, and this can be done via the Test Characteristic Curve [8].

2.2.2. *Conventional-EAP.* Some testing programs use the examinee's ability estimate, theta, to determine if the examinee passes or fails the test. Theta score are usually estimated based on sophisticated statistical methods which better estimated examinee's performance status. There are many ability estimation methods available, such as maximum likelihood estimation, maximum a posteriori estimation, Owen's Bayesian estimation and expected a posteriori estimation (EAP). The estimators from any of these methods could be used to determine an examinee's mastery status. However, EAP has been shown to have significantly less bias than the other methods [18]. Thus, it is common to use EAP to estimate examinees' ability levels.

In EAP estimation, the weights are the probabilities at the corresponding points of a discrete prior distribution. In some contexts of education, normal prior distributions for the points and weights are usually assumed to improve the accuracy of the numerical approximation of the integral [1]. In this Conventional-EAP procedure, the examinee's ability level is estimated by the EAP method after all items in the conventional test are administered. The EAP estimator is compared with the cut theta. If the EAP estimator is equal to or greater than the cut theta, then the examinee is classified as a master; on the other hand, if the EAP estimator is smaller than the cut theta, the examinee is classified as a non-master.

2.2.3. *Bayesian procedures.* There are two main components in the Bayesian procedure: the construction of the loss structure and the decision rule. These two components were implemented as described below together with an application of a Markov Chain Monte Carlo (MCMC) procedure using WinBUGS, as described in the Method section.

Generally speaking, a loss function specifies the total costs for each possible decision outcome. These costs usually incorporate all relevant psychological and social consequences associated with decisions. There are two kinds of loss functions used in this study: threshold loss function and linear loss function.

The specification of threshold loss functions follows [7] in which the expected losses associated with making a false positive error and a false negative error are specified by constants L and M, respectively (see Table 1).

Table 1. Threshold loss function defined for a fixed-length test

True theta level Decision made	Theta $<$ cut	Theta $\geq$ cut
Non-master	0	M
Master	L	0

Note: $-\infty < \theta < \infty$. Theta: examinee's true theta. L: the loss associated with a false positive error. M: the loss associated with a false negative error

Although the threshold loss function is simple and has been frequently used in the literature, it may be unrealistic in some situations [4]. A major criticism of threshold loss is that no matter how far the examinee's ability level is from the cut score, the loss is assumed to be equal.

To overcome the limitation of threshold loss, [12] proposed a linear loss function for fixed-length mastery tests, which assumes the loss to be a continuous function of the examinee's theta level. For a linear loss function, examinees with different theta levels have different loss functions. If an examinee is declared a master but his/her true theta level is below the cut score, then the linear loss function is a decreasing function of theta. On the other hand, if an examinee is declared a non-master but his/her true theta level is above the cut score, then the linear loss function is an increasing function of theta. The

expected losses associated with making a false positive error and a false negative error are specified as L $(\theta_{cut} - \theta)$ and M $(\theta - \theta_{cut})$. Table 2 provides the linear loss functions for four possible outcomes.

TABLE 2. Linear loss function defined for a fixed-length test

True theta level Decision made	Theta $<$ cut	Theta $\geq$ cut
Non-master	0	M $(\theta - \theta_{cut})$
Master	L $(\theta_{cut} - \theta)$	0

Note: $-\infty < \theta < \infty$. θ_{cut}: cut score. L and M are specified the same as in threshold loss functions.

Decision Rule

The decision rule is the crux of the Bayesian master/non-master classification procedure. However, in order to render valid classifications it must also account for the possibility of two kinds of errors: passing examinees who are true non-masters and failing examinees who are true masters [12]. The probabilities of these two classification errors are controlled through the use of loss functions. After administering a set of items, a Bayesian decision rule for a fixed-length mastery test is used to minimize the posterior expected loss associated with the two classification decisions. Suppose $P(\theta|X_n)$ represents the posterior distribution of theta after n items are administered;

$$\text{E [Loss (master, } P(\theta|X_n))] \text{ and E [Loss (non-master, } P(\theta|X_n))]$$

represent the posterior expected loss for making the mastery and non-mastery decisions, respectively. Based on [17], the examinee is classified as a master when

$$\text{E [Loss (master, } P(\theta|X_n))] < \text{E [Loss (non-master, } P(\theta|X_n))].$$

Otherwise, the examinee is classified as a non-master.

Replacing these terms in the previous equations, under the threshold loss function, the examinee is declared a master when

$$L * \int_{-\infty}^{\theta_{cut}} P(\theta|X_n)d\theta < M * \int_{\theta_{cut}}^{\infty} P(\theta|X_n)d\theta$$

And under the linear loss function, the examinee is declared a master when

$$L * \int_{-\infty}^{\theta_{cut}} (\theta_{cut} - \theta)P(\theta|X_n)d\theta < M * \int_{\theta_{cut}}^{\infty} (\theta - \theta_{cut})P(\theta|X_n)d\theta$$

The posterior distribution, $P(\theta|X_n)$, is the product of the prior and the likelihood. For this study, a vague prior for θ was used because it was desired that the prior distribution plays a minimal role in the posterior distribution and inferences. The prior was set for theta, and followed a normal distribution with mean equal to μ and the precision equal to τ (the precision is the inverse of the variance). The hyperprior on μ was a normal distribution with mean equal to 0 and the precision equal to 0.01. In order to constitute a conjugate prior distribution, the hyperprior on τ was a gamma distribution with parameters $(0.01, 0.01)$.

The likelihood is the product of the probabilities associated with each examinee's item responses. Suppose there are a total of N examinees responding to n items. The likelihood

function can be expressed as:

$$Likelihood = \prod_{j=1}^{N} \prod_{i=1}^{n} P_{ij}^{x_{ij}} \left(1 - P_{ij}^{1-x_{ij}}\right)$$

where x_{ij} is the item response on item i for a person j, $x_{ij} = 1$ or 0; P_{ij} is the probability of a correct response on item i for a person j based on the 3PL IRT model.

For both the threshold loss function and the linear loss function, three conditions were considered in this study: 2L = M, L = M, L = 2M, which represent the cost of making a false negative error was twice as serious as making a false positive error; the costs of making a false positive error and a false negative error were equal, and the cost of making a false positive error was twice as serious as making a false negative error, respectively. These ratios were utilized in the present study, as they are standard research conventions in previous studies [7,17,20].

3. **Method.** Monte Carlo simulations were utilized to compare the Bayesian procedure with different loss functions against two conventional procedures (Conventional-Proportion Correct and Conventional-EAP). The test lengths were set as 20, 40 and 60, which are common test lengths for most tests. For the item pool, the discrimination parameters were generated from a normal distribution, but with mean equal to 1 and standard deviation to 0.1. The difficulty parameters were generated from a normal distribution with mean equal to 0 and standard deviation equal to 2. The guessing parameters were set to be 0.15. These parameters were generated because the value is close to the real item pools. The cut score on the theta scale was equal to 0.4 for this study, which is the common cut theta level for many certification exams.

Sets of ability parameters for five thousand examinees were generated to fit a standard normal distribution (mean = 0 and standard deviation = 1.0). The responses of five thousand examinees to 20, 40, 60 items were then simulated. The probability of a correct response (P_{ij}) was compared with a random deviate (d_{ij}) which was drawn from a uniform distribution in the range $[0, 1]$. If $P_{ij} > d_{ij}$, the item was scored as correct (1); otherwise, the item was scored as incorrect (0). Both the simulation and data generation were conducted using the computer software program R.

4. **Results.** For each simulation, the outcomes of interest were (1) the percentages of correct classifications, (2) false positive error rates, (3) false negative error rates, and (4) phi correlations between the true classification status and observed classification status. In order to calculate these indices, the true masters and true non-masters needed to be defined first. The examinee's true theta level was compared with the cut score. If his/her true theta level was equal to or larger than the cut score, the examinee was truly a master; otherwise, the examinee was truly a non-master. Figures 1 to 4 present the results based on these four evaluation criteria.

Figures 1 to 4 showed that test length had an impact on the classification accuracy for both conventional methods and Bayesian procedures. When test length increased, the percentages of correct classifications and the corresponding phi correlations became higher while the false negative error rates and the false positive error rates became lower. For the percentages of correct classifications and phi correlations, the differences in these values between Bayesian procedures and the conventional procedures were quite small (approximately 5 to 8%). In addition, for the two types of loss functions-threshold loss functions and linear loss functions, there was no clear advantage for either loss function. For example, under the same simulation conditions, the Bayesian threshold L = 2M yielded the smallest false positive error rates but higher false negative error rates than

the Bayesian linear L = 2M. The threshold loss function seemed to control the false positive errors better, but in terms of controlling both false negative errors and false positive errors simultaneously, the linear loss function performed somewhat better than the threshold loss function.

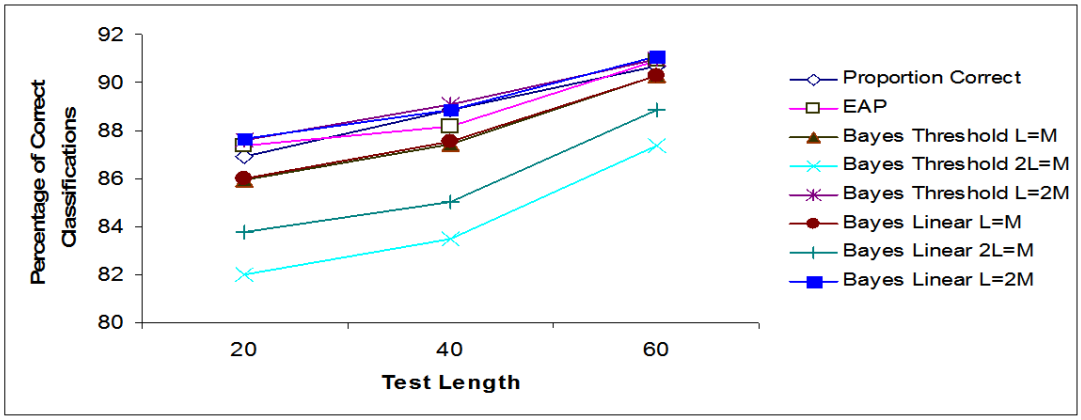

FIGURE 1. Percentages of correct classifications at each level of test length

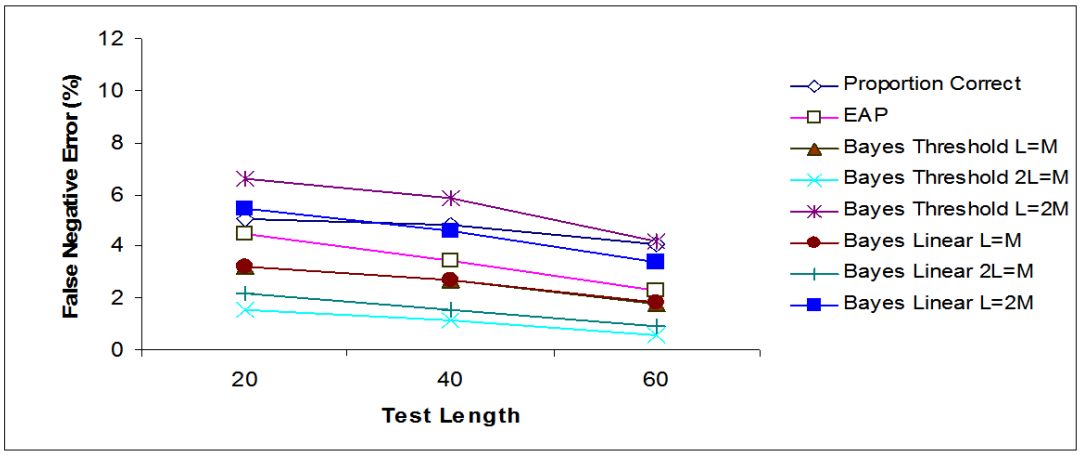

FIGURE 2. False negative error rates at each level of test length

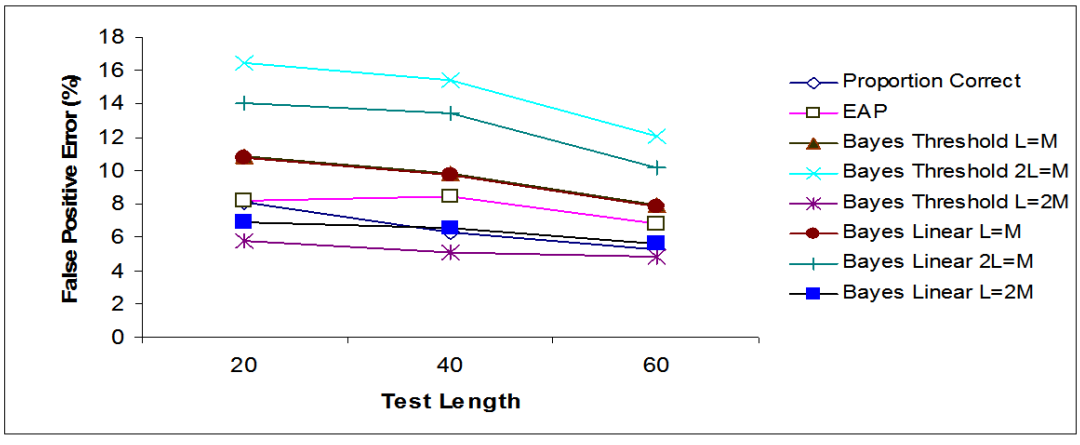

FIGURE 3. False positive error rates at each level of test length

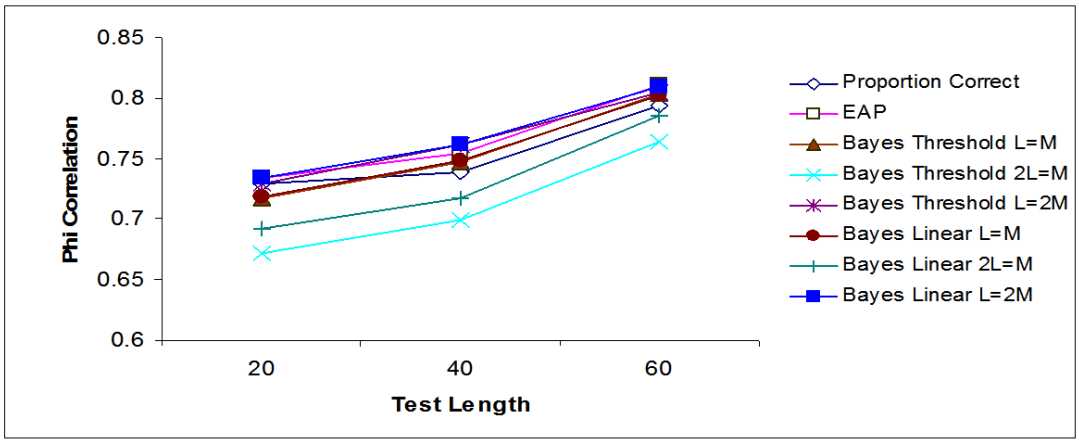

FIGURE 4. Phi correlations between true and predicted mastery status at each level of test length

5. **Conclusion and Discussion.** For the percentages of correct classifications and phi correlations, the ranges of these values between different procedures were quite small. However, based on an overall evaluation of the testing procedures, the loss function L = 2M yielded better results than the conventional-EAP and the conventional-Proportion correct procedures for all the test lengths in the high discrimination item pool and for the test length of 40 and 60 in the moderate discrimination item pool. In addition, by employing the appropriate loss parameters, the Bayesian procedure can effectively control false positive error rates and false negative error rates. This study considered two types of loss functions: threshold loss functions and linear loss functions. Generally, the threshold loss function is less desirable since it assumes a constant loss for all examinees. Some authors [12,16] noted that this assumption is probably unrealistic in some applications. It seems more realistic to assume that loss is an increasing function of theta for non-masters and a decreasing function of theta for masters [14,15]. Moreover, the threshold loss function is discontinuous at the cutoff point. The loss for correct and incorrect decisions should change smoothly rather than abruptly [13].

Theoretically, linear loss functions seem more desirable in real testing situations. However, the results from this study showed that the two loss functions performed similarly which could imply that using the linear loss or threshold loss function does not have much impact on making binary decisions, at least under the conditions manipulated in this study. Such a conclusion is consistent with the study conducted by [16].

In this study, the Bayesian procedure with loss function L = 2M (the cost of making a false positive error was twice as serious as making a false negative error) had higher accuracy indices than other Bayesian procedures in this study. This could be an artifact of the cut score used in this study. As stated previously, the cut score was set at a theta level of 0.4. With this cut score, approximately 65% of the simulees were nonmasters and 35% were masters since the population was simulated using a normal distribution. Thus, false positive errors can be considered less likely than false negative errors. It is quite likely that if the cut score were changed, the results would also change.

This study uses WinBUGS for the MCMC sampling. WinBUGS is well-established software used for Bayesian-related analyses. It is very easy to specify the models in WinBUGS and get the MCMC outputs. However, using WinBUGS to do the MCMC sampling was time-consuming and computer-intensive. For the conventional methods, it only took few minutes to estimate the classification errors in R. However, with the same computer, it took more than 4 hours to run the Bayesian decision-theoretic procedure with

a test length equal to 60 for 5000 examinees. In a real testing situation, there could be a very large number of students taking the exam. In addition, the score reports normally need to be finished within one to two weeks after receiving students' raw responses from the scanning center. Using the MCMC method in WinBUGS may not be realistic in operationally-based situations, since it could be difficult to meet the deadlines in some real testing situations. Until advances in hardware and programming are achieved, it might be more reasonable to use alternative numerical methods for making mastery decisions in such situations.

This study investigated some features that influence Bayesian decision-theoretic procedures in the context of fixed-format mastery testing using the 3PL IRT model. There were some limitations of this study. First, this study only considered one cut score. Different locations of cut scores on the theta scale should be considered to examine the impact of the two types of loss functions on the classification accuracy. Second, in this study, the b-parameters in the two simulated item pools were generated in a relatively broad range. Different types of item pools, such as uniform item pools, b-variable item pools, a- and b-variable item pools, a-, b-, c- variable item pools, should be investigated in the future to examine the influence of discrimination, difficulty and guessing parameters on the Bayesian procedure.

Third, this study only considered fixed-format mastery tests. It might be desirable to develop a variable-length format procedure to enhance the efficiency of test administration. Also, [13] indicated that an optimal situation for the sequential rules would be to choose an action (declaring pass, declaring fail, or continuing testing) that minimizes posterior expected loss at each stage of testing, using dynamic programming. This technique would consider the expected loss at the final stage of testing and then estimate backwards to the first stage of testing. In doing so, the action chosen would be optimal with regards to the entire sequential testing process. Currently, the implementation of this variable-length procedure would not be realistic using WinBUGS, since the processing speed is too slow.

Fourth, this study used four criteria (percentages of correct classifications, false positive error rates, false negative error rates and phi correlations between the true and observed classification status) to evaluate the results. Although these four indices are commonly used to evaluate the performance of testing procedures in mastery tests, some other criteria might reveal different trends. For example, [4] used the average of actual loss for all examinees (mean loss) to evaluate the performance of different test procedures.

Finally, this study only considered linear and threshold loss functions for the Bayesian decision-theoretic procedure. However, there are other types of loss structures that could be applied. For example, [10] presented procedures for specifying nonlinear loss functions for estimating examinees' ability levels in terms of cumulative distribution functions and using least square fitting techniques. This type of nonlinear loss function does not only reflect realistic situations but also can be incorporated with the standard normal distribution for the psychometric model.

Acknowledgement. This work is supported by the National Science Council (NSC 99-2511-S-656-001). The author acknowledged the helpful comments and suggestions made by Execute Editors, Anonymous Reviewers and Associate Editors that have greatly improved the quality of this paper.

REFERENCES

[1] R. D. Bock and R. J. Mislevy, Adaptive EAP estimation of ability in a microcomputer environment, *Applied Psychological Measurement*, vol.6, pp.431-444, 1982.

[2] H. C. Cho, K. S. Lee and M. S. Fadali, Online learning algorithm of dynamic Bayesian networks for nonstationary signal processing, *International Journal of Innovative Computing, Information and Control*, vol.5, no.4, pp.1027-1041, 2009.
[3] C. A. Glas and H. J. Vos, Adaptive mastery testing using the Rasch model and Bayesian sequential decision theory, *Research Report 98-15*, University of Twente, The Netherlands, 1998.
[4] C. A. W. Glas and H. J. Vos, Testlet-based adaptive mastery testing, *Computerized Testing Report 99-11*, Law School Admission Council, Newtown, PA, 2006.
[5] R. K. Hambleton, H. Swaminathan, J. Algina and D. B. Coulson, Criterion-referenced testing and measurement: A review of technical issues and developments, *Review of Educational Research*, vol.48, pp.1-47, 1978.
[6] R. Ji, X. Lang, H. Yao and Z. Zhang, Semantic sensitive region retrieval using keyword-integrated Bayesian reasoning, *International Journal of Innovative Computing, Information and Control*, vol.3, no.6(B), pp.1645-1656, 2007.
[7] C. Lewis and K. Sheehan, Using Bayesian decision theory to design a computerized mastery test, *Applied Psychological Measurement*, vol.14, pp.367-386, 1990.
[8] F. M. Lord, A broad-range tailored test of verbal ability, *Applied Psychological Measurement*, vol.1, no.1, pp.95-100, 1977.
[9] M. Fujisaki and D. Zhang, Bayesian analysis of compound poisson mixture model and its application to financial data, *International Journal of Innovative Computing, Information and Control*, vol.5, no.1, pp.109-117, 2009.
[10] M. R. Novick and D. V. Lindley, On the use of the cumulative distribution as a utility function in educational or employment selection, *Journal of Educational Measurement*, vol.15, no.3, pp.181-191, 1978.
[11] C.-W. Shen, Fault diagnosis for e-seal unreadability using learning Bayesian networks, *ICIC Express Letters*, vol.5, no.4(B), pp.1417-1421, 2011.
[12] W. J. van der Linden and G. J. Mellenbergh, Optimal cutting scores using a linear loss function, *Applied Psychological Measurement*, vol.1, pp.593-599, 1977.
[13] W. J. van der Linden, Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores, *Psychometrika*, vol.46, pp.257-274, 1981.
[14] W. J. van der Linden and J. H. Vos, A compensatory approach to optimal selection with mastery scores, *Psychometrika*, vol.61, pp.155-172, 1996.
[15] H. J. Vos, Applications of Bayesian decision theory to sequential mastery testing, *Research Report 97-06*, University of Twente, The Netherlands, 1997.
[16] H. J. Vos, Application of Bayesian decision theory to sequential mastery testing, *Journal of Educational and Behavioral Research*, vol.32, pp.403-433, 1999.
[17] H. J. Vos, A Bayesian's procedure in the context of sequential mastery testing, *Psicologica*, vol.21, pp.191-211, 2000.
[18] T. Wang, *The Precision of Ability Estimation Methods in Computerized Adaptive Testing*, The University of Iowa, 1995.