PSYCHOMETRIKA—VOL. 71, NO. 1, 121–137
MARCH 2006
DOI: 10.1007/s11336-003-1154-5

COMPUTERIZED ADAPTIVE TESTING UNDER NONPARAMETRIC IRT MODELS

XUELI XU

EDUCATIONAL TESTING SERVICE

JEFF DOUGLAS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Nonparametric item response models have been developed as alternatives to the relatively inflexible parametric item response models. An open question is whether it is possible and practical to administer computerized adaptive testing with nonparametric models. This paper explores the possibility of computerized adaptive testing when using nonparametric item response models. A central issue is that the derivatives of item characteristic Curves may not be estimated well, which eliminates the availability of the standard maximum Fisher information criterion. As alternatives, procedures based on Shannon entropy and Kullback–Leibler information are proposed. For a long test, these procedures, which do not require the derivatives of the item characteristic eurves, become equivalent to the maximum Fisher information criterion. A simulation study is conducted to study the behavior of these two procedures, compared with random item selection. The study shows that the procedures based on Shannon entropy and Kullback–Leibler information perform similarly in terms of root mean square error, and perform much better than random item selection. The study also shows that item exposure rates need to be addressed for these methods to be practical.

Key words: Shannon entropy, Kullback–Leibler information, nonparametric item response models, item response theory.

1. Introduction

Under the item response theory (IRT) framework, each examinee is indexed by a value of the latent variable θ. This latent variable is usually taken as undimensional. In addition, each item in the test is associated with an item characteristic curve (ICC), which specifies the probability of a correct response to the item as a nondecreasing function of θ. Nonparametric models have been developed to address the fact that the parametric models do not always fit the data adequately. Unlike parametric models with a limited number of parameters, nonparametric models allow much more flexibility to describe the probability of correct responses to items as a function of latent ability.

Whether nonparametric models should be used is a matter of assessing the trade-off between bias and variance. Nonparametric estimation can help reduce the bias that would result from fitting a misspecified parametric model. However, the cost will be increased variance at each point in the ICC estimate. This is because nonparametric estimation techniques rely primarily on local averaging, which essentially only utilizes a fraction of the observations for fitting the model at a single point. This is in contrast with parametric models which use more global criteria for determining the best estimate of a relatively small number of parameters. Nonparametric regression techniques, and nonparametric ICC estimation, are sometimes mistakenly thought of as techniques for use with small samples. In fact, nonparametric estimation is not suited for small samples, but can be quite beneficial in large sample situations. This is because the nonparametric

The authors would like to thank Hua Chang for his help in conducting this research.
Requests for reprints should be sent to Xueli Xu, Rosedale Road MS 02-T, Princeton, NJ 08541, USA.

nature of the fit allows for dramatically reducing bias, and when large enough samples are available the variance can also be controlled.

Methods of nonparametric ICC estimation include kernel smoothing with a selected scale for the latent trait (Ramsay, 1991; Douglas, 1997), monotone splines (Ramsay & Abrahamowicz, 1989), and penalized maximum likelihood estimation (Rossi,Wang, & Ramsay, 2002). The kernel smoothed ICC estimator is introduced in detail. Suppose N examinees are randomly sampled and take a test of length n. Let Y_{ij} be the binary response of examinee i to item j. The kernel smoothing estimator of $P_j(\theta)$ is a weighted average of the examinees' responses to item j,

$$\hat{P}_j(\theta) = \sum_{i=1}^{N} w_i(\theta - \theta_i) Y_{ij},$$

where the weights $w_i(\theta - \theta_i)$ are defined so that they are nonnegative and reach a maximum when $\theta = \theta_i$ and will approach or equal zero as $|\theta - \theta_i|$ increases. The weights $w_i(\theta - \theta_i)$ are defined by a kernel function $K(\cdot)$, with the properties mentioned above. Also, two additional conditions should be satisfied to make the kernel smoothing estimator meaningful. These two conditions are $w_i(\theta - \theta_i) \geq 0$ and $\sum_i w_i(\theta - \theta_i) = 1$. A popular choice of weights are given by Nadaraya and Watson (Nadaraya, 1964; Watson, 1964):

$$w_i(\theta - \theta_i) = \frac{K\left(\frac{\theta - \theta_i}{h}\right)}{\sum_i K\left(\frac{\theta - \theta_i}{h}\right)},$$

where $K(\cdot)$ denotes the kernel function, and h refers to a bandwidth which governs the degree of smoothing. The kernel smoothing estimator of $P_j(\theta)$ is consistent when θ_i can be estimated without error. However, the latent trait values of θ_i are not observable. The Nadaraya–Watson weights can still be used after substituting the true θ_i with $\hat{\theta}_i$. A common and appropriate way to construct $\hat{\theta}_i$ is to rank the summed scores and transform to the corresponding quantile of the chosen distribution $F(\theta)$. This leads to the kernel smoothing estimator

$$\hat{P}_j(\theta) = \frac{\sum_{i=1}^{N} K\left(\frac{\theta - \hat{\theta}_i}{h}\right) Y_{ij}}{\sum_{i=1}^{N} K\left(\frac{\theta - \hat{\theta}_i}{h}\right)},$$

proposed by Ramsay (1991) and implemented in *TestGraf* (Ramsay, 2000). The consistency of this estimator was proved by Douglas (1997). The use of summed scores in this way is justified for long exams by the asymptotic results of Douglas (1997) that $\hat{\theta}$ is consistent for θ. Large deviation theorems are given as well as convergence rates. A finite test length justification is seen in Grayson's (1988) work on the monotone likelihood ratio property of summed scores for general monotone IRT models with binary items. This result implies the stochastic ordering of the latent trait for different summed scores.

In kernel smoothing, the bandwidth h is used to control the balance between the bias and variance of estimation. At this point, there is no theorem on an optimal bandwidth for ICC estimation. However, we can use results from simpler models where the covariate is measured without error as a guideline. For example, Ramsay (1991) suggested that $h = N^{-1/5}$ works well when using a Gaussian kernel. In nonparametric regression problems with constant error variance, which is not satisfied in ICC estimation with binary response variables, it can be shown that the bandwidth that minimizes the mean squared error is a multiple of $N^{-1/5}$ (Eubank, 1988). This can be expected to hold in the binary case as well, but the optimal constant by which $N^{-1/5}$ is multiplied would be more difficult to determine. In our application, this bandwidth appeared to recover the true curve well for the given experimental conditions.

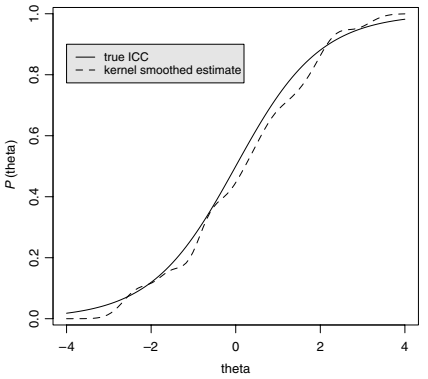

FIGURE 1.
A true ICC and its estimate.

One application of item response models is test assembly. Here we consider one aim of educational testing, which is to estimate the latent ability of an examinee on the studied domain. Traditional paper and pencil tests present a sample of examinees with the same set of items. A possible consequence is that neither high ability groups nor low ability groups may simultaneously be measured with high reliability. One remedy for this is adaptive testing. For each examinee, adaptive testing selects future items adapted to the current ability estimate. With the development of item response models and the increasing power of computers, computerized adaptive testing (CAT) has become feasible. Though there are many psychometrically challenging issues associated with CAT (van der Linden & Glas, 2000), this paper only focuses on item selection algorithms for CAT.

The most popular method used in CAT practice is the maximum information criterion (MIC). For a latent variable θ, let $L_n(\theta)$ be the log-likelihood function of θ after n items have been administered. Denote the Fisher information of θ by

$$I_n(\theta) = E([\partial L_n(\theta)/\partial\theta]^2).$$

Let $\hat{\theta}_n$ be the maximum likelihood estimator of the true θ. Then it is well known that $\hat{\theta}_n$ has limiting distribution $N(\theta, 1/I_n(\theta))$ as $n \to \infty$. This property of $\hat{\theta}_n$ motivated Birnbaum's (1968) development of MIC. In the MIC, the next item is selected from the items remaining in the item bank to maximize the Fisher information at the present estimate $\hat{\theta}_n$.

Notice that $I_j(\theta)$ involves taking the derivative of $\hat{P}_j(\theta)$. This can be very problematic when using nonparametrically estimated ICCs. Even methods that provide good estimates of the ICCs may not yield acceptable derivative estimates. For instance, consider kernel smoothed ICC estimates. The derivative of the estimate will be negative at some values of θ. This can be seen in Figures 1 and 2. Even monotone splines will result in quite jagged derivative estimates. For this reason, the MIC cannot be reliably used with an item bank of nonparametrically estimated ICCs. Consequently, new methods that do not involve derivatives will be better choices. In the following sections, a procedure based on Shannon entropy and a procedure based on Kullback–Leibler information are introduced and a simulation study is conducted.

2. Kullback–Leibler Procedure

This procedure is inspired by Chang and Ying's (1996) "global information criterion" (GIC). Generally, the Kullback–Leibler (K–L) information is a discrepancy measure between two probability distributions. Let $K(f(x), g(x))$ be the K–L information, where $f(x)$ and $g(x)$

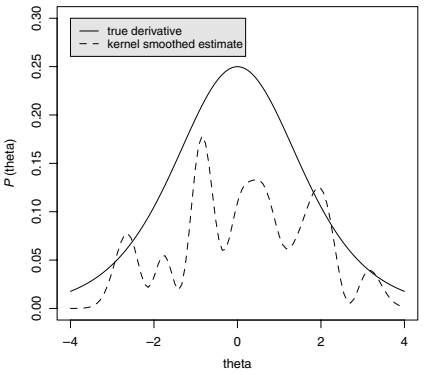

FIGURE 2.
First derivative of a true ICC and its estimate.

are two probability density functions with $f(x)$ being the true density. Then the K–L information is defined as

$$K(f, g) = \int f(x) \ln \frac{f(x)}{g(x)} \mu(dx).$$

In the expression above, μ denotes the dominating measure for densities f and g. In item response models where X is discrete, μ will be counting measure, and the integral sign is replaced with the summation sign. In the context of IRT, $f(x)$ and $g(x)$ are replaced with the likelihood functions evaluated at different values of θ that are induced by a candidate item. Chang and Ying (1996) defined the K–L information for the jth item as

$$K_j(\hat{\theta}_n, \theta) = E\left[\ln \frac{L(\hat{\theta}_n|Y_j)}{L(\theta|Y_j)}\right],$$

where $\hat{\theta}_n$ is the maximum likelihood estimate of θ after n items have been administered, and the expectation is taken with respect to the future item response Y_j for a candidate item j. This can be written as

$$K_j(\hat{\theta}_n, \theta) = P_j(\hat{\theta}_n) \log\left[\frac{P_j(\hat{\theta}_n)}{P_j(\theta)}\right] + [1 - P_j(\hat{\theta}_n)] \log\left[\frac{1 - P_j(\hat{\theta}_n)}{1 - P_j(\theta)}\right].$$

An index of global information is defined by integrating $K_j(\hat{\theta}_n, \theta)$ over an interval centered at $\hat{\theta}_n$. Denote this integral by

$$G_j(\hat{\theta}_n) = \int_{\hat{\theta}_n - \delta_n}^{\hat{\theta}_n + \delta_n} K_j(\hat{\theta}_n, \theta)\, d\theta,$$

where δ_n is an approximate margin-of-error in the estimation of θ that decreases at a rate of $n^{-1/2}$. Then the GIC is to select the (n+1)st item among the remaining items in the item bank that maximizes $G_j(\hat{\theta}_n)$, where j indexes the items remaining in the item bank.

Chang and Ying (1996) noted that the MIC and GIC are asymptotically equivalent. Thus, as n becomes quite large, the MIC and GIC will select nearly identical (if not identical) items. They also demonstrate with simulation studies that the GIC is superior to the MIC for a short test.

Notice that $K_j(\hat{\theta}_n, \theta)$ does not involve any derivatives of the ICCs. An implication of this is that the GIC may easily be used with nonparametrically estimated item response models as well as with parametric item response models.

3. Shannon Entropy Procedure

Shannon entropy (Shannon, 1948; Cover & Thomas, 1991) was first proposed as a measure of the complexity of discrete probability distributions. It can be extended to the case where the random variable is continuous. Let $f(x)$ be the density function of a random variable x, then the Shannon entropy is defined as

$$SH(f) = E[-\ln f(x)].$$

For a discrete random variable, the Shannon entropy will reach its maximum value when all the points in sample space are of equal probability. It reaches a minimum of zero when all the probability is concentrated at a single point. In the case of continuous probability distributions, Shannon entropy will be small when the distribution of the random variable is concentrated in a smaller interval, and becomes even smaller when the distribution vanishes to a single point. This property of Shannon entropy serves our purpose for CAT with nonparametric item response models. The basic aim of sequential item selection is to make the posterior distribution of the ability parameter become as concentrated as possible at the true value. By selecting an item that minimizes the expected Shannon entropy of the posterior distribution we can achieve this. This idea was introduced by DeGroot (1962) and was applied as the "Shannon entropy procedure" in Tatsuoka and Ferguson (2003) and Tatsuoka (2002) for their partially ordered set model for cognitive diagnosis, and was also studied for application in cognitive diagnosis by Xu, Chang, and Douglas (2003).

Let j index the items remaining in the item bank after n items have been administered, and let $\pi_{n,j}$ denote the posterior distribution after administering n items and this future item j. Depending on the value of the item response Y_j, we may compute the expected Shannon entropy of $\pi_{n,j}$, given the current posterior distribution π_n. This is given by

$$ESH(\pi_{n,j}) = \sum_{y_j=0,1} SH(\pi_{n,j}) p(Y_j = y_j \mid y_1, \ldots, y_n),$$

where the expectation is taken over the density of Y_j given the current posterior distribution of θ. The Shannon entropy procedure is to select the next item from the remaining items in the item bank to minimize $ESH(\pi_{n,j})$. Note that an alternative criterion would be to select the next item to minimize the expected posterior variance. In some related work, the authors have found that selecting items in this manner is not as efficient as the criterion based on Shannon entropy.

Let $I_n(\theta_0)$ be the Fisher information at the true value of θ, after the first n items have been administered. The following theorem shows the asymptotic equivalence of $I_n(\theta_0)$ and $SH(\pi_{n,j})$.

Theorem 1. *Given the regularity conditions stated in Assumptions* 1 *through* 5 *of Appendix* A, $SH(\pi_n) - (-\ln I_n(\theta_0)^{1/2})$ *converges to* $1/2 + \ln\sqrt{2\pi}$ *as* $n \to \infty$, *with probability equal to* 1.

Theorem 1 suggests that the Shannon entropy procedure and the Fisher information computed at true ability carry precisely the same information as n becomes large. A proof of Theorem 1 is given in Appendix A along with a heuristic argument that the MIC and the Shannon entropy procedure should result in the same items in a fairly large item bank.

4. Simulation Study

The performance of the Shannon entropy procedure and the K–L procedure as well as random item selection are compared in a simulation study. It involves calibrating an item bank with kernel smoothed ICC estimates when data are simulated from a 2PL model.

4.1. Simulation Study

4.1.1. Item bank description. An item bank was generated by first simulating responses of 1000 subjects with $N(0, 1)$ distributed abilities to 500 items from a two-parameter logistic model with ICCs given by

$$P_j(\theta) = \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]}.$$

For the 500 items, the discrimination parameters were drawn from a uniform distribution on (0.75, 2.5) and the difficulty parameters were drawn from a standard normal distribution. Then the item bank was constructed based on kernel smoothed estimates of the ICCs using a Gaussian kernel with a bandwidth of 0.25, which is $1000^{-1/5}$. When fitting the kernel smoothed ICCs, the distribution of the latent ability was assumed to be uniform on (0, 1). Despite the fact that a $N(0, 1)$ ability distribution was used to generate the item responses, this uniform (0, 1) scale may be used when items are estimated nonparametrically, -because the functional form of the ICCs is not constrained to have a particular form after a change of variables. In evaluating CAT methods, we retain this scale for the latent ability.

4.1.2. Simulation Design. Three approaches, the Shannon entropy procedure, the K–L procedure, and random item selection, were compared with respect to root mean squared error(RMSE),bias and exposure rates across the simulations. They are defined in the following. $RMSE_S$ is the RMSE of latent ability estimates given a certain length of subtest, while $RMSE_\theta$ is the RMSE of latent ability estimates given certain value of θ,

$$RMSE_S = \sqrt{\sum_{i=1}^{N} (\hat{\theta}_{i,S} - \theta_i)^2/N},$$

$$RMSE_\theta = \sqrt{\sum_{i:\theta_i=\theta} (\hat{\theta}_i - \theta)^2/\#I(\theta_i = \theta)}.$$

The bias, given a certain stage of the test, and the bias given a certain value of θ are denoted by $bias_S$ and $bias_\theta$, respectively,

$$bias_S = \sum_{i=1}^{N} (\hat{\theta}_{i,S} - \theta_i)/N,$$

$$bias_\theta = \sum_{i:\theta_i=\theta} (\hat{\theta}_i - \theta)/\#I(\theta_i = \theta).$$

In each simulation, the test length was fixed at 50 items. However, the performance of these procedures could be evaluated at any length of subtest. For each procedure a total of 10,000 examinees was generated from a uniform (0, 1) ability distribution. Each examinee was given exactly the same five items to begin the test, then the maximum likelihood estimate was obtained. The next item was selected sequentially according to different strategies until the final test length was reached. After each stage of administering an item, the maximum likelihood ability estimate was obtained.

4.1.3. Results. The upper part of Figure 3 shows the RMSE and bias comparisons for these three methods as a function of the length of the exam. The two proposed methods outperform random item selection in both aspects, and they have very similar results. The same pattern is

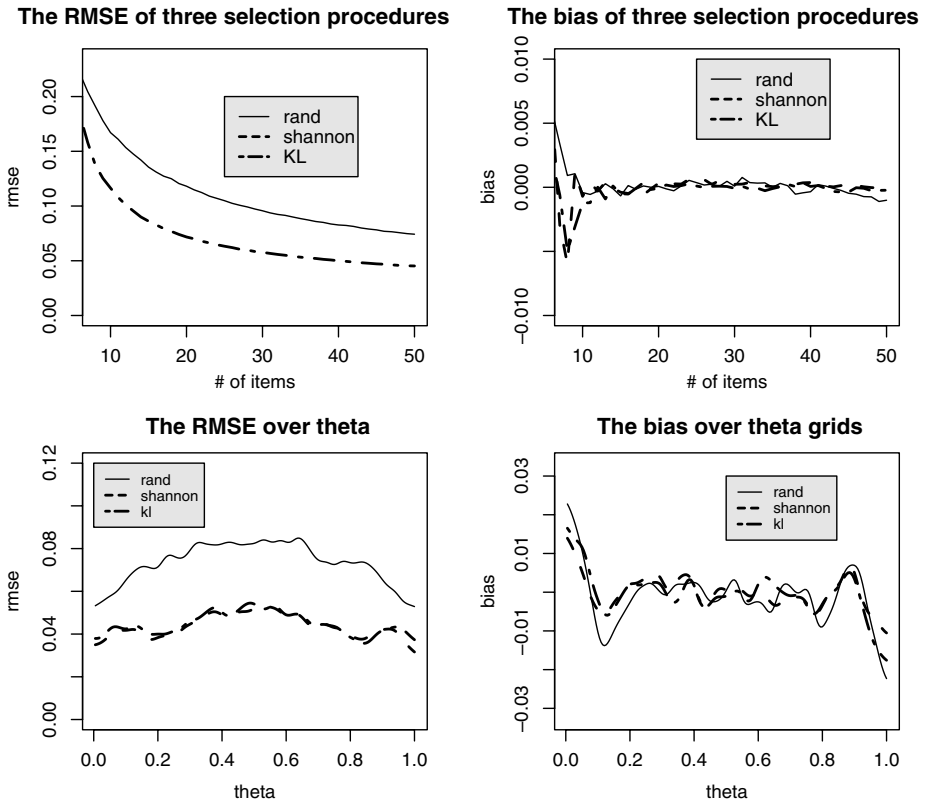

FIGURE 3.
Root mean square error and bias comparison among approaches.

shown in the lower part of Figure 3, which presents an RMSE and bias comparison as a function of θ. Let's look at one aspect of the results. Table 1 gives the RMSE at different lengths of the subtest. After 40 items, both the K–L procedure and the Shannon entropy procedure result in sufficient accuracy in estimating the latent ability.

In addition, both methods result in high proportions of items that are not frequently used. Figures 4 through 7 present the exposure rate of items. Figure 4 summarizes the item exposure rates for the whole range ability under three approaches. If the ability can be divided into three categories: low (0.01–0.33), medium (0.34–0.66), and high (0.67–0.99), then Figures 5 through 7 present us with the item exposure rate for these three ability intervals. Compared with the random item selection, both the K–L procedure and the Shannon entropy procedure result in a relatively

TABLE 1.
The average root mean square error for the three methods.

	Number of items			
Selection rule	20	30	40	50
Random	0.118	0.096	0.083	0.074
K–L	0.072	0.057	0.049	0.045
Shannon	0.071	0.056	0.049	0.044

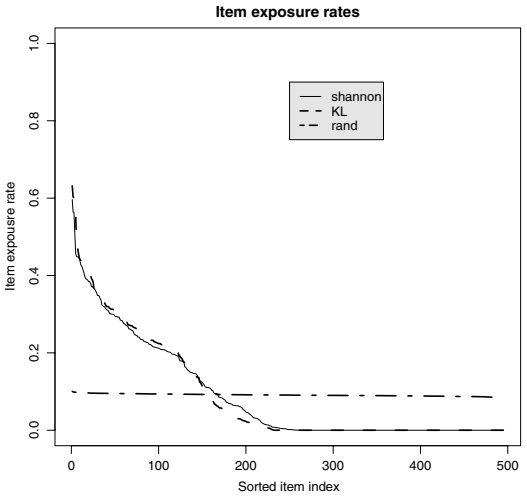

FIGURE 4.
Item exposure rates.

high proportion of low-exposure items and high-exposure items. These numbers are not good enough if we want to make full use of the item bank. Further research on how to use these two methods, while balancing item exposure control, will be needed.

Concerning computing time, the K–L procedure is much faster than the Shannon entropy procedure due to the calculation of the posterior distribution required for Shannon entropy. For example, for a test fixed at 50 items, it took only 6 seconds to finish the test for 100 examinees, while it took 231 seconds to finish the test using the Shannon entropy procedure.

5. Discussion

This paper concerns the possibility of computerized adaptive testing under nonparametric IRT models. Two algorithms, the K–L procedure and the Shannon entropy procedure were proposed as candidate algorithms. Neither procedure involves computing of the derivatives of ICCs, which is difficult for nonparametrically estimated ICCs. Both procedures lead to fast convergence of ability estimation. The K–L distance procedure is noted to be equivalent to the MIC (Chang & Ying, 1996) and Shannon entropy is proved to be equivalent to the MIC in this paper. The simulation study shows that these two methods perform similarly in terms of the RMSE of the latent ability estimation. The simulation study also shows that after 30 items have been administered, their selection of the very next item is identical 85.4% of the time. In addition, both methods lead to high exposure rates for some items, and to a high proportion of items that are not frequently used. This should be addressed in applications.

6. Appendix

6.1. Asymptotic Equivalence of the Shannon Entropy Procedure and the MIC

In this appendix it will be shown that the MIC and the Shannon entropy procedure can be expected to result in selecting the same item in a finite item bank when the test length n becomes large provided that the derivatives of ICCs could be computed for the information function used

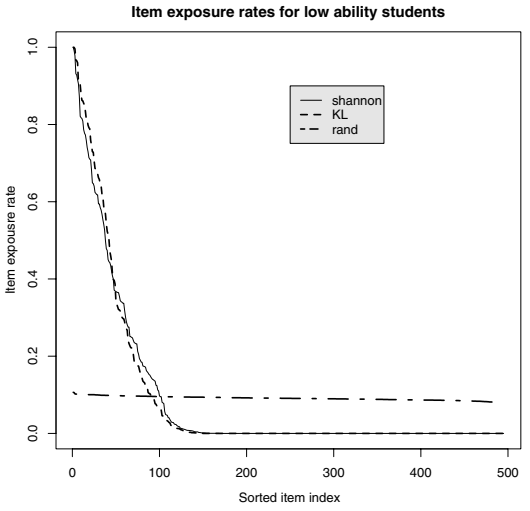

FIGURE 5.
Item exposure rates.

in the MIC. First, an important property of the Shannon entropy of the posterior distribution is stated as Theorem 1. Theorem 1 makes use of the notation and assumptions used in Chang and Stout (1993), in their proof of the asymptotic posterior normality of the latent ability.
Basic notation:
$P_j(\theta)$: the ICC for item j given θ.
θ_0: the true latent ability. Let Y_j be a binary random variable, it is assumed that Y_j has the density $P_j(\theta_0)^{y_j}[1 - P_j(\theta_0)]^{1-y_j},\ y_j = 0, 1.$
$\hat{\theta}_n$: the maximum likelihood estimator (MLE) of θ.
$I_n(\theta)$: the Fisher information of θ accumulated after n items have been administered.
$\hat{\sigma}_n$: defined as $[I_n(\hat{\theta}_n)]^{-1/2}$.

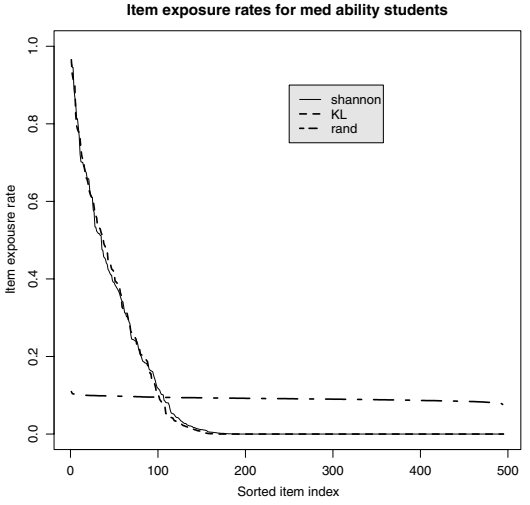

FIGURE 6.
Item exposure rates.

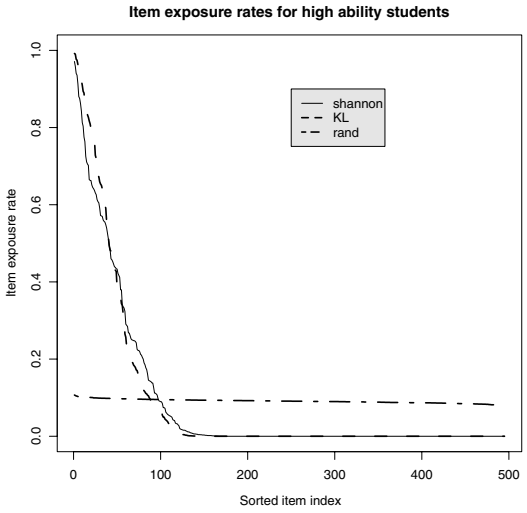

FIGURE 7.
Item exposure rates.

$L_n(\theta)$: the log-likelihood function of latent ability θ after n items have been administered. Sometimes we write it as $\ln P(Y_1, \ldots, Y_n|\theta)$.
$\lambda_j(\theta)$: the logit function of item j, $\lambda_j(\theta) = \ln(P_j(\theta)/(1 - P_j(\theta)))$.
$\pi_0(\theta)$: the prior density function of θ.
$\pi_n(\theta)$: the posterior density function of θ. It has the form

$$\pi_n(\theta) = \frac{P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)}{\int P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)\, d\theta}.$$

Assumption 1. *Let $\theta \in \Theta$, where Θ is $(-\infty, \infty)$ or a bounded or unbounded interval in $(-\infty, \infty)$. Let the prior density $\pi_0(\theta)$ be Lipschitz continuous and positive at θ_0, where θ_0 is assumed to be the true value of θ. Further, assume that $E[\ln(\pi_0(\theta))] < \infty$.*

Assumption 2. *$P_j(\theta)$ is twice continuously differentiable and the first and second derivatives are bounded in absolute value uniformly with respect to θ in some closed interval $|\theta - \theta_0| < \delta$.*

Assumption 3. *For every fixed $\theta \neq \theta_0$, assume, for some given $c(\theta) > 0$,*

$$\limsup n^{-1} E_{\theta_0}(L_n(\theta) - L_n(\theta_0)) \leq -c(\theta),$$

and

$$\sup_j |\lambda_j(\theta)| < \infty.$$

Assumption 4. *The information function for each item has a first derivative, and $\lambda_j(\theta)$ has second and third derivatives. All these derivatives are bounded in absolute value uniformly in j and in $|\theta - \theta_0| < \delta$.*

Assumption 5.

$$\liminf_{n\to\infty} \frac{I_n(\theta_0)}{n} > c(\theta_0) > 0,$$

where $c(\theta_0)$ is a constant.

The proof of Theorem 1 uses three lemmas that are similar to those in Chang and Stout (1993). These lemmas are somewhat stronger and require slight modifications of Chang and Stout's proof. These details are not given here, but are available upon request.

Lemma 1. *Suppose the assumptions hold. For any sequence* $\delta_n > c\sqrt{\ln n/n}$, *for some constant* c, *there exists a* $k(\delta_n) > 0$ *such that*

$$P_{\theta_0}\{\limsup_{n\to\infty, |\theta-\theta_0|>\delta_n} n^{-1}[L_n(\theta) - L_n(\theta_0)] < -k(\delta_n)\} = 1.$$

Lemma 2. *Suppose the assumptions hold. Then*

$$L_n(\theta) - L_n(\hat{\theta}_n) = (\theta - \hat{\theta}_n)^2 L_n''(\hat{\theta}_n^*)/2 = -\frac{(\theta - \hat{\theta}_n)^2}{2\hat{\sigma}_n^2}(1 - R_n),$$

where $\hat{\theta}_n^*$ *is a point between* θ *and* $\hat{\theta}_n$, *and* R_n *is defined as* $R_n = 1 + \hat{\sigma}_n^2 L_n''(\hat{\theta}_n^*)$. *Let* $\epsilon_n = 2\sqrt{2\ln n/nC'}$, *then with probability 1,*

$$\sup_{|\theta-\theta_0|<\delta_n} |R_n| < \epsilon_n,$$

where $C' = c^2/8\zeta_{\lambda_2}$ *and* ζ_{λ_2} *is the bound for the second derivative of* λ_j, $\delta_n = C\epsilon_n$ *for some constant* C.

Lemma 3. (Chang and Stout, 1993). *Given the assumptions above,*

$$(P(Y_1, \ldots, Y_n|\hat{\theta}_n)\hat{\sigma}_n)^{-1} \int P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)\, d\theta \to \sqrt{2\pi}\pi_0(\theta_0)$$

with probability 1.

The proof of Theorem 1 uses techniques similar to those used in the proof of the posterior normality of the latent trait in Chang and Stout (1993).

Proof of Theorem 1. Let $A_n = (P(Y_1, \ldots, Y_n|\hat{\theta}_n)\hat{\sigma}_n)^{-1}$ and let $P(Y_1, \ldots, Y_n) = \int P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)\, d\theta$, then

$$\begin{aligned} Sh(\pi_n) - (-\ln I_n(\theta_0)^{1/2}) &= -\int \pi_n(\theta)\ln\pi_n(\theta)d\theta - (-\ln I_n(\theta_0)^{1/2}) \\ &= -\int A_n \frac{P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)}{A_n P(Y_1, \ldots, Y_n)} \ln\left[\frac{A_n P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)}{A_n P(Y_1, \ldots, Y_n)}\right] d\theta \\ &\quad -(-\ln I_n(\theta_0)^{1/2}) \\ &= -\frac{G1}{A_n P(Y_1, \ldots, Y_n)} + G2 - (-\ln I_n(\theta_0)^{1/2}), \end{aligned}$$

where $G1 = \int A_n P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)\ln[A_n P(Y_1, \ldots, Y_n|\theta)\pi_0(\theta)]\, d\theta$ and $G2 = \ln[A_n P(Y_1, \ldots, Y_n)]$. From Lemma 3, we know that $G2 \to \ln[\sqrt{2\pi}\pi_0(\theta_0)]$ with probability 1. For part G1, we need to consider two different subsets of Θ, $|\theta - \theta_0| > \delta_n$ and $|\theta - \theta_0| \le \delta_n$, where δ_n is

defined as in Lemma 2,

$$
\begin{aligned}
G1 &= \int A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)\ln[A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)]\,d\theta \\
&= \int_{|\theta-\theta_0|>\delta_n} A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)\ln[A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)]\,d\theta \\
&\quad + \int_{|\theta-\theta_0|\le\delta_n} A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)\ln[A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)]\,d\theta.
\end{aligned}
$$

After plugging in A_n, the first part of the integral becomes

$$
\begin{aligned}
&\int_{|\theta-\theta_0|>\delta_n} A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)\ln[A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)]\,d\theta \\
&= \int_{|\theta-\theta_0|>\delta_n} \exp(L_n(\theta)-L_n(\theta_0)+L_n(\theta_0)-L_n(\hat\theta_n))\hat\sigma_n^{-1}\pi_0(\theta)\ln[\exp(L_n(\theta) \\
&\quad -L_n(\hat\theta_n))\hat\sigma_n^{-1}\pi_0(\theta)]\,d\theta \\
&\le \int_{|\theta-\theta_0|>\delta_n} \exp(L_n(\theta)-L_n(\theta_0))\pi_0(\theta)\hat\sigma_n^{-1}\ln[\exp(L_n(\theta)-L_n(\hat\theta_n))\hat\sigma_n^{-1}\pi_0(\theta)]\,d\theta \\
&= \int_{|\theta-\theta_0|>\delta_n} \exp(L_n(\theta)-L_n(\theta_0))\pi_0(\theta)\hat\sigma_n^{-1}(L_n(\theta)-L_n(\theta_0))\,d\theta \\
&\quad + \int_{|\theta-\theta_0|>\delta_n} \exp(L_n(\theta)-L_n(\theta_0))\pi_0(\theta)\hat\sigma_n^{-1}(L_n(\theta_0)-L_n(\hat\theta_n))\,d\theta \\
&\quad + \int_{|\theta-\theta_0|>\delta_n} \exp(L_n(\theta)-L_n(\theta_0))\pi_0(\theta)\hat\sigma_n^{-1}\ln\hat\sigma_n^{-1}\,d\theta \\
&\quad + \int_{|\theta-\theta_0|>\delta_n} \exp(L_n(\theta)-L_n(\theta_0))\pi_0(\theta)\hat\sigma_n^{-1}\ln(\pi_0(\theta))\,d\theta.
\end{aligned}
$$

The inequality sign is due to the fact that $\exp(L_n(\theta_0)-L_n(\hat\theta_n))\le 1$. By Lemma 1, the first part of the integral of $G1$ converges to 0 with probability 1,

$$
\begin{aligned}
&\int_{|\theta-\theta_0|>\delta_n} A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)\ln[A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)]\,d\theta \\
&\le \exp(-nk(\delta_n))I_n(\hat\theta_n)^{1/2}(-nk(\delta_n)) + (L_n(\theta_0)-L_n(\theta_n))\ \exp(-nk(\delta_n))I_n(\hat\theta_n)^{1/2} \\
&\quad + \exp(-nk(\delta_n))I_n(\hat\theta_n)^{1/2}\ln(I(\hat\theta_n)^{1/2}) \\
&\quad + \exp(-nk(\delta_n))I_n(\hat\theta_n)^{1/2}\int_{|\theta-\theta_0|>\delta_n}\pi_0(\theta)\ln(\pi_0(\theta))\,d\theta \\
&\to 0.
\end{aligned}
$$

Plugging in A_n and applying Lemma 2, the second part of the integral of G_2 becomes

$$
\begin{aligned}
&\int_{|\theta-\theta_0|\le\delta_n} A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)\ln[A_n P(Y_1,\ldots,Y_n|\theta)\pi_0(\theta)]\,d\theta \\
&\quad = \int_{|\theta-\theta_0|\le\delta_n} \exp(L_n(\theta)-L_n(\theta_n))\sigma_n^{-1}\pi_0(\theta)\ln\left[(L_n(\theta)-L_n(\theta_n))\sigma_n^{-1}\pi_0(\theta)\right]d\theta
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{|\theta-\theta_0|\le\delta_n} \exp\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)\right]\hat\sigma_n^{-1}\pi_0(\theta)\\
&\quad\times\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)+\ln\hat\sigma_n^{-1}+\ln\pi_0(\theta)\right]d\theta\\
&= \int_{|\theta-\theta_0|\le\delta_n} \exp\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)\right]\hat\sigma_n^{-1}\pi_0(\theta)\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)\right]d\theta\\
&\quad+\int_{|\theta-\theta_0|\le\delta_n}\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)\right]\hat\sigma_n^{-1}\pi_0(\theta)\ln\hat\sigma_n^{-1}d\theta\\
&\quad+\int_{|\theta-\theta_0|\le\delta_n} \exp\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)\right]\hat\sigma_n^{-1}\pi_0(\theta)\ln\pi_0(\theta)\,d\theta\\
&= D_1+D_2+D_3
\end{aligned}
$$

We analyze the limiting behavior of G_2 by using similar techniques as those in Walker (1969) and Chang and Stout (1993). Assumption 1 states that $\pi_0(\theta)$ is Lipschitz continuous and positive at $\theta=\theta_0$. Hence, for $|\theta-\theta_0|\le\delta_n$, we have

$$1-K\delta_n \le \inf_{|\theta-\theta_0|\le\delta_n}\frac{\pi_0(\theta)}{\pi_0(\theta_0)} \le \sup_{|\theta-\theta_0|\le\delta_n}\frac{\pi_0(\theta)}{\pi_0(\theta_0)} \le 1+K\delta_n,$$

where $K>0$ is a constant. Then

$$(1+K\delta_n)D_{11} < \pi_0(\theta_0)^{-1}D_1 < (1-K\delta_n)D_{11},$$

where

$$D_{11}=\int_{|\theta-\theta_0|\le\delta_n} \exp\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)\right]\hat\sigma_n^{-1}\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-R_n)\right]d\theta.$$

By Lemma 2, we know for $|\theta-\theta_0|<\delta_n$, we have $\sup|R_n|<\epsilon_n$. It is seen that D_{11} is bounded above by

$$\int_{|\theta-\theta_0|\le\delta_n} \exp\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1+\epsilon_n)\right]\hat\sigma_n^{-1}\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-\epsilon_n)\right]d\theta,$$

and is bounded below by

$$\int_{|\theta-\theta_0|\le\delta_n} \exp\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1-\epsilon_n)\right]\hat\sigma_n^{-1}\left[-\frac{(\theta-\hat\theta_n)^2}{2\hat\sigma_n^2}(1+\epsilon_n)\right]d\theta.$$

These upper and lower bounds can be viewed as multiples of the second central moments of normal variables over an interval, provided that $\hat\theta_n$ is the mean and $\hat\sigma_n^2$ is the variance. Since the $\hat\theta_n$ is strongly consistent and $\hat\sigma_n$ converges to 0 almost surely, these second central moments over an interval will converge to the second central moments of normal variables. Thus,

$$P_{\theta_0}\{\lim_{n\to\infty} I[-1/2\sqrt{2\pi}(1-\epsilon_n)^{-3/2}(1+\epsilon_n) < D_{11} < -1/2\sqrt{2\pi}(1+\epsilon_n)^{-3/2}(1-\epsilon_n)]=1\}=1.$$

Since $\epsilon_n \to 0$ and $\delta_n \to 0$, we will find that D_1 converges to $-(1/2)\pi_0(\theta_0)\sqrt{2\pi}$ with probability 1 as $n \to \infty$.

Considering D_3, for $|\theta - \theta_0| < \delta_n$, we can have, from Assumption 1, that $\pi_0(\theta)\ln(\pi_0(\theta))$ is in an interval

$$(\pi_0(\theta_0)\ln\pi_0(\theta_0) - h_1(\delta_n), \quad \pi_0(\theta_0)\ln\pi_0(\theta_0) + h_2(\delta_n)),$$

where $h_1(\delta_n) > 0$ and $h_2(\delta_n) > 0$ go to 0 as $\delta_n \to 0$. In fact, since

$$1 - K\delta_n \leq \frac{\pi_0(\theta)}{\pi_0(\theta_0)} \leq 1 + K\delta_n,$$

and

$$\ln(1 - K\delta_n) \leq \ln\frac{\pi_0(\theta)}{\pi_0(\theta_0)} \leq \ln(1 + K\delta_n),$$

$h_1(\delta_n)$ and $h_2(\delta_n)$ could be any absolute values of multiplication of these bounds as long as $h_1(\delta_n) \leq h_2(\delta_n)$. Thus,

$$D_3 < \pi_0(\theta_0)\ln\pi_0(\theta_0)D_{31} + h_2(\delta_n)D_{31},$$

where D_{31} is

$$D_{31} = \int_{|\theta-\theta_0|\leq\delta_n} \exp\left[-\frac{(\theta - \hat{\theta}_n)^2}{2\hat{\sigma}_n^2}(1 - R_n)\right]\hat{\sigma}_n^{-1}d\theta.$$

By Lemma 2,

$$P_{\theta_0}\left\{\lim_{n\to\infty} I[\sqrt{2\pi}(1+\epsilon_n)^{-1/2} < D_{31} < \sqrt{2\pi}(1-\epsilon_n)^{-1/2}] = 1\right\} = 1.$$

Hence

$$(\pi_0(\theta_0)\ln\pi_0(\theta_0) + h_2(\delta_n))D_{31} \to \sqrt{2\pi}\pi_0(\theta_0)\ln\pi_0(\theta_0)$$

for large enough n, with probability 1, as $\epsilon_n \to 0$. By using the same argument,

$$D_3 > (\pi_0(\theta_0)\ln\pi_0(\theta_0) - h_1(\delta_n))D_{31},$$

and the right part of this inequality will converge to $\sqrt{2\pi}\pi_0(\theta_0)\ln\pi_0(\theta_0)$ with probability 1, as $\epsilon \to 0$. Therefore, D_3 converges to $\sqrt{2\pi}\pi_0(\theta_0)\ln\pi_0(\theta_0)$ for large n, with probability 1.

Now we move to the part D_2. Let

$$B = -\sqrt{2\pi}\pi_0(\theta_0)\ln I_n(\theta_0)^{1/2},$$

and let

$$B' = -\sqrt{2\pi}\ln I_n(\theta_0)^{1/2}\int_{|\theta-\theta_0|\leq\delta_n}\frac{1}{\sqrt{2\pi}\hat{\sigma}_n}\exp\left[-\frac{(\theta-\hat{\theta}_n)^2}{2\hat{\sigma}_n^2}(1-R_n)\right]\pi_0(\theta)\,d\theta.$$

First, we want to show the behavior of the integral. By Assumption 1, the Lipschitz continuity of $\pi_0(\theta)$ and Lemma 2, for $|\theta - \theta_n| \leq \delta_n$, the integral is bounded above by

$$(1 + K\delta_n)(1-\epsilon_n)^{-1/2}\pi_0(\theta_0)\left\{\Phi\left[\left(\frac{\theta_0 - \hat{\theta}_n + \delta_n}{\hat{\sigma}_n}\right)(1+\epsilon_n)^{1/2}\right]\right.$$
$$\left. - \Phi\left[\left(\frac{\theta_0 - \hat{\theta}_n - \delta}{\hat{\sigma}_n}\right)(1+\epsilon_n)^{1/2}\right]\right\},$$

and is bounded below by

$$(1-K\delta_n)(1+\epsilon_n)^{-1/2}\pi_0(\theta_0)\left\{\Phi\left[\left(\frac{\theta_0-\hat{\theta}_n+\delta_n}{\hat{\sigma}_n}\right)(1-\epsilon_n)^{1/2}\right]\right.$$
$$\left.-\Phi\left[\left(\frac{\theta_0-\hat{\theta}_n-\delta}{\hat{\sigma}_n}\right)(1-\epsilon_n)^{1/2}\right]\right\},$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Both the brackets in the upper and lower bounds converge to unity with probability 1 as $n\to\infty$. Hence the integral part of $B^{'}$ converges to $\pi_0(\theta_0)$ with probability 1. Next, we show that

$$-D_2-B^{'}\to 0 \quad \text{almost surely},$$

and

$$B^{'}-B\to 0 \quad \text{almost surely}.$$

From Assumptions 4 and 5 and the strong consistency of θ_n under regularity conditions, it follows that $\ln I_n(\hat{\theta}_n)^{1/2}-\ln I_n(\theta_0)^{1/2}$ is equal to $I_n^{'}(\theta_0)/I_n(\theta_0)\,(\hat{\theta}_n-\theta^*)$, for some θ^* between θ_0 and $\hat{\theta}_n$. By the regularity conditions, $I_n^{'}(\theta^*)/I_n(\theta^*)$ is continuous and bounded in a neighborhood of θ_0. Thus, the strong consistency of $\hat{\theta}_n$ implies that $\ln I_n(\hat{\theta}_n)^{1/2}-\ln I_n(\theta_0^{1/2})$ will converge to 0 with probability 1. Then,

$$-D_2-B' = -\sqrt{2\pi}[\ln I_n(\hat{\theta}_n)^{1/2}-\ln I_n(\theta_0)^{1/2}]$$
$$\times\int_{|\theta-\theta_0|\le\delta_n}\frac{1}{\sqrt{2\pi}\hat{\sigma}_n}\exp\left[-\frac{(\theta-\hat{\theta}_n)^2}{\hat{\sigma}_n^2}(1-R_n)\right]\pi_0(\theta)\,d\theta.$$

Since the integral in the expression above will go to $\pi_0(\theta_0)$ almost surely, we have that $-D_2-B^{'}\to 0$ with probability 1.

Next we will prove that $B^{'}-B$ converges to 0 almost surely. Notice that

$$B^{'}-B=-\sqrt{2\pi}\ln I_n(\theta_0)\left[\int_{|\theta-\theta_0|<\delta_n}\frac{1}{\sqrt{2\pi}\hat{\sigma}_n}\exp\left[-\frac{(\theta-\hat{\theta}_n)^2}{2\hat{\sigma_n}^2}(1-R_n)\right]\pi_0(\theta)d\theta-\pi_0(\theta_0)\right].$$

By applying the upper bounds and lower bounds for the integral and the Lipschitz continuity of $\pi_0(\theta)$, $B^{'}-B$ is bounded below by

$$-\sqrt{2\pi}\ln I_n(\theta_0)\pi_0(\theta_0)((1+K\delta_n)(1-\epsilon_n)^{-1/2}-1)$$
$$\times\left\{\Phi\left[\left(\frac{\theta_0-\hat{\theta}_n+\delta_n}{\hat{\sigma}_n}\right)(1+\epsilon_n)^{1/2}\right]-\Phi\left[\left(\frac{\theta_0-\hat{\theta}_n-\delta_n}{\hat{\sigma}_n}\right)(1+\epsilon_n)^{1/2}\right]\right\},$$

and is bounded above by

$$-\sqrt{2\pi}\ln I_n(\theta_0)\pi_0(\theta_0)((1-K\delta_n)(1+\epsilon_n)^{-1/2}-1)$$
$$\times\left\{\Phi\left[\left(\frac{\theta_0-\hat{\theta}_n+\delta_n}{\hat{\sigma}_n}\right)(1-\epsilon_n)^{1/2}\right]-\Phi\left[\left(\frac{\theta_0-\hat{\theta}_n-\delta_n}{\hat{\sigma}_n}\right)(1-\epsilon_n)^{1/2}\right]\right\}$$

The brackets in the upper and lower bounds will converge to 1 for large n with probability 1. By L'Hôpital's theorem, $((1-K\delta_n)(1+\epsilon_n)^{-1/2}-1)/1/\ln n$ converges to 0. Similarly, $((1+K\delta_n)(1-\epsilon_n)^{-1/2}-1)/1/\ln n$ converges to 0. Since $\ln(I_n(\theta_0))$ is $O(\ln n)$, then we have that both the upper and lower bounds converge to 0 almost surely. Thus $B^{'}-B$ will converge to 0.

Finally, combining all the results above,

$$
\begin{aligned}
SH(\pi_n) - (-\ln I_n(\theta_0)^{1/2}) \quad &= \\
&- \int \pi_n(\theta) \ln \pi_n(\theta) d\theta - (-\ln I_n(\theta_0)^{1/2}) \\
&= -\frac{G1}{A_n P(Y_1, \ldots, Y_n)} + G2 - (-\ln I_n(\theta_0)^{1/2}) \\
&= G2 - \frac{D_1 + D_3}{A_n P(Y_1, \ldots, Y_n)} - \frac{D_2}{A_n P(Y_1, \ldots, Y_n)} - \frac{\sqrt{2\pi}\pi_0(\theta_0)}{\sqrt{2\pi}\pi_0(\theta_0)} \\
&\quad (-\ln I_n(\theta_0)^{1/2}) \\
&\to^{\text{a.s.}} \ln[\sqrt{2\pi}\pi_0(\theta_0)] - \frac{-1/2\sqrt{2\pi}\pi_0(\theta_0) + \sqrt{2\pi}\pi_0(\theta_0)\ln\pi_0(\theta_0)}{\sqrt{2\pi}\pi_0(\theta_0)} + 0 \\
&= 1/2 + \ln\sqrt{2\pi}.
\end{aligned}
$$

Theorem 1 shows that for any sequence of items, the difference between the Shannon entropy and $-\ln I_n(\theta_0)^{1/2}$ converges to a constant with probability 1, as $n \to \infty$. Because of the strong convergence of $\ln I_n(\hat{\theta}_n)$, combining Theorem 1,

$$SH(\pi_n) - (-\ln I_n(\hat{\theta}_n)^{1/2}) \to 1/2 + \ln\sqrt{2\pi} \quad \text{almost surely.}$$

The only way to guarantee the difference as a constant is to optimize $SH(\pi_n, Y_n)$ and $-\ln I_n(\hat{\theta}_n)$ in the same direction. Therefore minimizing the Shannon entropy means to maximize the $I_n(\hat{\theta}_n)$. Hence, to minimize the $ESH_j(\pi_{n,j})$ is to minimize $SH(\pi_{n,j})$, which is meant to maximize the $I_{n+1}(\hat{\theta}_n)$.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479) Reading, MA: Addison-Wesley.

Chang, H.H., & Stout, W.F. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

Chang, H.H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229.

Cover, T.M., & Thomas, J.A. (1991). *Elements of information theory.* New York: Wiley.

DeGroot M.H. (1962). Uncertainty, information and sequential experiments. *Annals of Mathematical Statistics*, *33*, 404–419.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7–28.

Eubank, R.L. (1988). *Spline smoothing and nonparametric regression.* New York, Marcel Dekker.

Grayson, D.A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392.

Nadaraya, E.A. (1964). On estimating regression. *Probability Theory and its Applications*, *9*, 141–142.

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630.

Ramsay, J.O. (2000). TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data [Computer Program]. Montreal: McGill University.

Ramsay, J.O., & Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, *84*, 906–915.

Rossi, N., Wang, X., & Ramsay, J.O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *27*, 291–317.

Shannon, C.E. (1948). A mathematical theory of communication, *Bell Systems Techical Journal*, *27*, 379–423, 623–656.

Tatsuoka, C. (2002). Data analytic methods for latent parially ordered classification models. *Journal of the Royal Statistical Society*, *Series C, 51*, 337–350.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on patially ordered sets. *Journal of Royal Statistical Society*, Series B, *65*, 143–158.

van der Linden, W.J., & Glas, C.A.W. (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

Walker, A.M. (1969). On the asymptotic behavior of posterior distributions. *Journal of the Royal Statistical Society, Series B*, *31*, 80–88.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Series A*, *26*, 359–372.

Xu X., Chang, H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Presented at the Annual Meeting of the National Council of Measurement in Education, Chicago, April 2003.

Manuscript received 23 DEC 2003
Final version received 30 OCT 2004
Published Online Date: 16 MAR 2006