A Simulation Study to Compare CAT Strategies for Cognitive Diagnosis

Xueli Xu
Department of Statistics,University of Illinois
Hua-Hua Chang
Department of Educational Psychology,University of Texas
Jeff Douglas
Department of Statistics,University of Illinois

April 18, 2003

Abstract

This paper demonstrates the performance of two possible CAT selection strategies for cognitive diagnosis. One is based on Shannon entropy and the other is based on Kullback-Leibler information. The performances of these two test construction methods are compared with random item selection. The cognitive diagnosis model used in this study is a simplified version of the Fusion model. Item banks are constructed for the purpose of simulation. The major result is that the Shannon entropy procedure outperforms the procedure based on Kullback-Leibler information in terms of correct classification rates. However, Kullback-Leibler has slightly smaller item exposure rates than the Shannon entropy procedure. This study shows that the Shannon entropy procedure is a promising CAT criterion, but modification might be required to control the exposure rate.

1 Introduction

1.1 Cognitive Diagnosis Models

Cognitive diagnosis models can be used to detect the presence or absence of specific skills required for educational exams. Unlike item response models that summarize the examinee's skill with a single broadly defined latent trait, cognitive diagnosis models utilize a high-dimensional latent vector with components that specify the presence or absence of specific skills or abilities. For instance, consider an algebra test. Under the IRT scheme, the objective is to measure the general ability of algebra. However, in cognitive diagnosis the aim might be to assess a multitude of cognitive skills, such as factoring, laws of exponents and manipulating fractions.

To date, at least fourteen distinct cognitive diagnosis models have appeared in the literature (Hartz, 2002; Roussos, 1994). In this paper we introduce two of them, the NIDA model and the Fusion model, and use the Fusion model in the simulation.

To describe these models, consider an item loading matrix with N items and M attributes. This matrix is called a Q matrix and was first introduced by K. Tatsouka (1984). $Q = \{Q_{jk}\}, j = 1, ..., N, k = 1, ..., M$, with $Q_{jk} = 1$ if item j requires attribute k to perform the task, with $Q_{jk} = 0$ otherwise. Each row of Q is a list of the cognitive attributes that an examinee needs to have mastered in order to give a correct response to the item. Table 1 gives an illustration of a Q matrix for an algebra assessment. This exam consists of 10 items, with each item requiring up to 3 attributes of the specified attributes.

Table 1: The Q matrix for a hypothetical algebra assessment

item - skill	factoring	laws of exponents	manipulating fractions
1	1	1	1
2	1	0	0
3	0	1	0
4	0	0	1
5	1	0	1
6	1	1	0
7	0	1	1
8	0	0	1
9	1	0	1
10	0	1	1

For instance, item 1 loads on all three attributes, while item 8 only loads on skill 3, i.e., 'manipulating fractions'.

In both models, the latent variable is a vector of 0's and 1's, indicating the presence and absence of cognitive skills. We denote it as $\boldsymbol{\alpha_i'} = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iM})$. Let X_{ij} be a matrix of 0's and 1's, representing the observed performance of examinee i on item j. The objective is to make inferences about the latent variable α_{ik}, or make inferences about the relationship between these attributes and test items.

The NIDA Model

The Noisy Inputs, Deterministic 'And' gate model (NIDA) was recently discussed by Maris (1999). The notation of Junker and Sijtsma (2001) for the NIDA model will be used in the following introduction. In this model, the latent response variable $\eta_{ijk} = 1$ or 0 is defined, indicating whether examinee i correctly applies the required attribute k to item j. The latent response variables are connected to $\boldsymbol{\alpha_i}$ through the

following two probabilities:

$$s_k = P(\eta_{ijk} = 0 | \alpha_{ik} = 1, Q_{jk} = 1)$$

and

$$g_k = P(\eta_{ijk} = 1 | \alpha_{ik} = 0, Q_{jk} = 1)$$

and $P(\eta_{ijk} = 1 | Q_{jk} = 0) = 1$ by definition. We refer to s_k and g_k as 'slip' and 'guess' probabilities, respectively. Observed item responses are deterministically related to the latent response variables through the conjunctive model $X_{ij} = \prod_k (\eta_{ijk})^{Q_{jk}}$. Thus, assuming local independence of the latent responses within an item, the IRF for item j is:

$$\begin{aligned} P(X_{ij} = 1 | \boldsymbol{\alpha_i}, \boldsymbol{s}, \boldsymbol{g}) &= \prod_k P(\eta_{ijk} = 1 | \alpha_{ik}, Q_{jk}, s_k, g_k) \\ &= \prod_k [(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}]^{Q_{jk}} \\ &= P_j(\boldsymbol{\alpha_i}) \end{aligned}$$

Assuming local independence among items as well as independence among examinees, the joint likelihood of NIDA model is:

$$\begin{aligned} L(\boldsymbol{s}, \boldsymbol{g}; \boldsymbol{\alpha}) &= \prod_i \prod_j P_j(\boldsymbol{\alpha_i})^{x_{ij}} (1 - P_j(\boldsymbol{\alpha_i})^{1-x_{ij}}) \\ &= \prod_i \prod_j \{\prod_k [(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}]^{Q_{jk}}\}^{x_{ij}} \{1 - \prod_k [(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}]^{Q_{jk}}\}^{1-x_{ij}} \end{aligned}$$

The Fusion Model

The Fusion model is an extension of the NIDA model in the way that it takes the incompleteness of the specified attributes of the Q matrix into account, as well as the way it lets attribute guessing and slipping parameters be item specific. Let η_{ijk} be as defined in the NIDA model, denoting whether examinee i correctly applies a required attribute k to item j. Parameters π_{jk} and r_{jk} are defined in a similar way as s_k and g_k, respectively, but allow variation across the items:

$$\pi_{jk} = P(\eta_{ijk} = 1|\alpha_{ik} = 1)$$

and

$$r_{jk} = P(\eta_{ijk} = 1|\alpha_{ik} = 0)$$

This useful generalization creates a source of non-identifiability that is rectified by a more parsimonious reparameterization (Hartz,2002). Let

$$\pi_j^* = \prod_k \pi_{jk}^{Q_{jk}}$$

and

$$r_{jk}^* = P(\eta_{ijk} = 1|\alpha_{ik} = 0)/P(\eta_{ijk} = 1|\alpha_{ik} = 1) = r_{jk}/\pi_{jk}$$

Another difference between the NIDA model and the Fusion models relates to the 'completeness' of Q. In the Fusion model, a completeness index for item j is denoted by c_j. Then the IRF for a single item is:

$$P(X_{ij}|\boldsymbol{\alpha}, \pi_j^*, r_{jk}^*, c_j, \theta) = \pi_j^* \prod_k (r_{jk}^{*(1-\alpha_{ik})Q_{jk}}) P_{c_j}(\theta_i)$$

where $P_{c_j}(\theta_i)$ is a Rasch IRF with difficulty parameter c_j. The latent variable θ_i essentially serves as a random effect, and represents attributes needed for the test but not accounted for in Q. In this study, we dropped the completeness part of the Fusion model and assumed that the Q matrix is complete. This simplified form of the Fusion model is the model used for our simulation study.

1.2 Computerized Adaptive Testing(CAT)

Maximizing Fisher information is a major criterion for item selection in computer adaptive testing. This criterion, however, can no longer be used in cognitive diagnosis because of mathematical restrictions. As we know, Fisher information is the expectation of the second derivative of log-likelihood equation with respect to the latent trait. In cognitive diagnosis models, the attribute vector $\boldsymbol{\alpha}$ consists of 0's and 1's. The derivatives with respect to $\boldsymbol{\alpha}$ are not defined, so that we cannot use Fisher information. Nevertheless, the idea of 'maximizing information' is still a primary concern. In this paper, two other appropriately defined information functions are used: Kullback-Leibler information and Shannon entropy.

2 METHODS

2.1 Kullback-Leibler information

Let $K(f, g)$ represent Kullback-Leibler (K-L) information for distinguishing between probability density functions f and g when f is true. It is defined as:

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

In this notation, μ is a dominating measure for densities f and g, and hence the integral sign above is replaced with a summation sign when X is a discrete variable, as in the case of item response variables. Generally, Kullback-Leibler information is a measure of the 'distance' between the two likelihoods. The larger this information, the easier to differentiate the two likelihoods (Lehmann and Casella, 1998). Kullback-Leibler information has been used as a criterion for item selection in IRT (Chang and Ying, 1996) to deal with the problem of large item parameter estimation error that can occur in the beginning of the test.

In our context, we use it as a criterion to select the next item that gives the largest K-L distance between our current estimate of $\boldsymbol{\alpha}'$ and other competing values of $\boldsymbol{\alpha}$. Let $\boldsymbol{\alpha}_{\boldsymbol{i}}' = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iM})$ denote the latent attribute vector, consisting of 0's and 1's, with respect to M cognitive attributes. Suppose at the present stage n-1 items have been selected, and denote this set of items as S_{n-1}. Let W represent the whole item bank; then define $R_n = W \backslash S_{n-1}$, the remaining items in the item bank. The Kullback-Leibler index for the jth item given the current estimate $\hat{\boldsymbol{\alpha}}$ is defined as the

sum of the Kullback-Leibler distances between $\hat{\boldsymbol{\alpha}}$ and possible candidate attribute vectors $\boldsymbol{\alpha_c}$ generated by the jth item.

$$K_j(\hat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^M}[\sum_{x=0}^{1} \log(\frac{P(X_j = x|\hat{\boldsymbol{\alpha}})}{P(X_j = x|\boldsymbol{\alpha_c})})P(X_j = x|\hat{\boldsymbol{\alpha}})].$$

The inner sum represents the K-L information for the distribution of $j-th$ item depending on attribute vectors $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha_c}$ when $\hat{\boldsymbol{\alpha}}$ is regarded as a true value. Then our criterion for selecting the nth item in the test is to select $j \varepsilon R_n$ such that $K_j(\hat{\boldsymbol{\alpha}})$ is maximized. For a thorough theoretical treatment of Kullback-Leibler information in item selection see Tatsouka and Ferguson (2003).

2.2 Shannon entropy procedure

Shannon entropy

The other method we consider for item selection is based on Shannon entropy (Cover, and Thomas, 1991). We begin with the properties of Shannon entropy. Let Ω denote the sample space of a random variable Y. For simplicity, it is assumed that Ω contains K possible values, $y_1, y_2, \ldots, y_K$. Let $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ be the probability vector of these K values, such that $\pi_i > 0$, $i = 1, ..., K$, and $\sum_{i=1}^{K} \pi_i = 1$. Thus,

$$P(Y = y_i) = \pi_i \qquad for \qquad i = 1, \ldots, K$$

. Denote the Shannon entropy of $\boldsymbol{\pi}$ by

$$Sh(\boldsymbol{\pi}) = \sum_{i=1}^{K} \pi_i \log(1/\pi_i).$$

The Shannon entropy $Sh(\boldsymbol{\pi})$ is thus a function of a random variable's probability distributions. It has the following properties:

1). Shannon entropy is a nonnegative, concave function that reaches its maximum when $\pi_1 = \pi_2 = ... = \pi_K = 1/K$. For instance, suppose there are only two points y_1, y_2 in Ω, with probabilities p and $q = 1 - p$. The Shannon entropy is $Sh(p) = -p \log p - (1-p)\log(1-p)$, $0 < p < 1$. The relationship between $Sh(p)$ and p can be seen clearly from Figure 1. When $p = 1/2$, $Sh(p)$ reaches its maximum. When $p = 0$ or 1, $Sh(p) = 0$.

2). Another property is that the more concentrated the distribution is, the lower the Shannon entropy. For example, suppose that there are five points in Ω, and their corresponding probabilities are $\pi_1, \pi_2, \pi_3, \pi_4, \pi_5$, with $\sum_{i=1}^{5} \pi_i = 1$. By a simple calculation, we have

$$Sh(\boldsymbol{\pi}) = \begin{cases} 0 & if \quad \pi_1 = 1 \\ 0.693 & if \quad \pi_1 = \pi_2 = \frac{1}{2} \\ 1.098 & if \quad \pi_1 = \pi_2 = \pi_3 = \frac{1}{3} \\ 1.386 & if \quad \pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{1}{4} \\ 1.609 & if \quad \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \frac{1}{5} \end{cases}$$

We observe that the Shannon entropy of $\boldsymbol{\pi}$ is smaller when a few points account for most of the probability. In our example, the distribution that loads entirely on one point has much smaller Shannon entropy than the distribution that allocates probability equally to five points. This notion underlies the Shannon entropy procedure for item selection. Suppose we are fitting a cognitive diagnosis model in a Bayesian context in which a prior distribution for the attribute vector $\boldsymbol{\alpha}$ is specified.

Let $\boldsymbol{\pi}$ be the posterior distribution of the attribute vector $\boldsymbol{\alpha}$ after administering n items. By further selecting items that minimize the Shannon entropy of $\boldsymbol{\pi}$, then we arrive at a posterior distribution that places all of its mass at the true attribute vector.

Shannon entropy procedure

In this section we present a formal description of the Shannon entropy procedure for item selection. More technical details about the Shannon entropy in general can be found in DeGroot (1962) and Tatsouka (2002). The aim of CAT in the context of cognitive diagnosis is to make the posterior probability of the true attribute pattern $\boldsymbol{\alpha}$ quickly approach 1 by sequentially selecting items, which have calibrated item parameters, that minimize the expected Shannon entropy of the posterior distribution of the attribute vector. Assume that there are M attributes, resulting in 2^M attribute patterns. Denote the prior probabilities of the 2^M patterns by $\boldsymbol{\pi_0}(\boldsymbol{\alpha_c}), c = 1, ..., 2^M$, and $\sum_{c=1}^{2^M} \pi_0(\boldsymbol{\alpha_c}) = 1$. The posterior distribution of a candidate pattern $\boldsymbol{\alpha}_c$ after n-1 items have been administered is, using Bayes formula

$$\pi_{n-1}(\boldsymbol{\alpha_c}) \propto \pi_0(\boldsymbol{\alpha_c}) \prod_{j=1}^{n-1} P_j^{x_j}(\boldsymbol{\alpha_c})[1 - P_j(\boldsymbol{\alpha}_c)]^{(1-x_j)},$$

where $P_j(\boldsymbol{\alpha_c})$ denotes the IRF of item response X_j given attribute pattern $\boldsymbol{\alpha_c}$. Then the Shannon entropy of the posterior distribution of the attribute pattern is:

$$E_{n-1}(\boldsymbol{\pi_{n-1}}) = \sum_{c=1}^{2^M} \pi_{n-1}(\boldsymbol{\alpha_c}) \log(1/\pi_{n-1}(\boldsymbol{\alpha_c})).$$

Once again, let W represent the whole item bank, and define $R_n = W \backslash S_{n-1}$, the

remaining items in the item bank. For item X_j in R_n, the expected Shannon entropy after administering X_j, where the expectation is computed under the current posterior distribution $\boldsymbol{\pi}_{n-1}$ is

$$\begin{aligned} Sh(\boldsymbol{\pi_n}, X_j) &= \sum_{x=0}^{1} E_n(\boldsymbol{\pi}_n | X_j = x) P[X_j = x | \boldsymbol{\pi}_{n-1}] \\ &= \sum_{x=0}^{1} \{E_n(\boldsymbol{\pi}_n | X_j = x)(\sum_{c=1}^{2^M} P_j^x(\boldsymbol{\alpha_c})[1 - P_j(\boldsymbol{\alpha_c})]^{1-x} \pi_{n-1}(\boldsymbol{\alpha_c}))\} \end{aligned}$$

Then our criterion for the n-th item is to select $j \varepsilon R_n$ such that $Sh(\boldsymbol{\pi_n}, X_j)$ is maximized. Optimality properties of this procedure have been studied by DeGroot (1962) and Tatsouka and Ferguson (2003).

3 Simulation study design

3.1 item bank construction

Parameters for simulation were selected based on several criteria. First, we wanted to consider a wide range of Fusion model parameters to enhance the generalizability of the results. Second, we required that the chosen parameters result in data sets yielding classical item statistics in a realistic range. Finally, the number of attributes measure per item was chosen to have a distribution capable of providing adequate measurement for each attribute. Two item banks were constructed for the purpose of simulation. In item bank 1, the Q matrix was generated to have varying complexity across the items, which can be seen in Figure 2. The elements of the r^* matrix were

generated from a uniform (0.3,0.8) distribution. If $Q_{jk} = 0$, then $r^*_{jk} = 1$ was set equal to 1. The π^*s were generated from a uniform (0.3,0.8) distribution. In order to make the r^* matrix and the π^* matrix realistic, the items were calibrated using BILOG. Only the items having realistic IRT item parameters were kept in the item bank. Here we required that guessing parameters be smaller than 0.4, discrimination parameters ranging from 0.3 to 1.5, and difficulty parameters ranging from -2.5 to 2.5. By these criteria, nearly 20% of the items were deleted from the item bank. Finally, item bank 1 ended up with 400 items, requiring up to 5 attributes. In item bank 2, the Q matrix was generated in such a way that the proportion of the attributes loading on each item is shown as in Figure 3. The r^* matrix was also generated from a uniform (0.3,0.8) distribution, but the π^*s were generated from a uniform (0.3,1) distribution. Following the same BILOG calibration procedure as that in constructing item bank 1, item bank 2 ended up with 420 items, measuring up to 8 attributes. The distribution of r^* for each attribute in these two item banks are shown in Figure 4 and Figure 5, respectively.

3.2 simulation design of CAT

Two different test lengths, 30 items and 50 items, were examined in both item banks. Under each testing environment, 2500 examinees were generated uniformly from the space of possible attribute patterns. For example, for item bank 1, examinees were generated from patterns 1 to 32 with equal probability, while for item bank 2, ex-

aminees were generated from patterns 1 to 256 with equal probability. This uniform distribution on the attribute patterns was chosen to minimize any benefit that an informative prior distribution on the attribute patterns would have on correct classification rates. In that regard, the results given here can be viewed as conservative. Within each testing environment, each examinee started from the same fixed set of 5 items that were randomly chosen from the item bank, and then the next item was selected according to different selection rules, K-L information, Shannon entropy procedure or random selection (as a baseline). The starting items may differ across examinees, but they were the same across the three different methods with the same examinee. The performances of these three methods were compared in terms of individual attribute classification rates, whole attribute pattern recognition, and item exposure rates.

4 Results

The correct classification rates are shown in Tables 2 to 5. Table 2 and Table 3 show the classification rates for five individual attributes and for the whole pattern. Table 4 and Table 5 show the classification rates for eight individual attributes and for the whole pattern. The last two columns of Table 2 through Table 5 display the comparison to random selection and K-L, respectively, in terms of the whole pattern recognition rate. For example, the numbers in the last column of Table 2 represent the proportion correctly classified relative to K-L classification. Specifically, 0.739

means that random selection is 26.1% less accurate than KL, while 1.536 means the Shannon entropy procedure is 53.6% more accurate than KL.

Table 2: Correct classification rates for 30-item-5-attribute test (2500 trials)

	1	2	3	4	5	whole pattern	whole pattern ratio to 'random'	whole pattern ratio to 'K-L'
random	0.782	0.750	0.750	0.781	0.744	0.291	1.000	0.739
Shannon	0.918	0.909	0.839	0.898	0.870	0.605	2.079	1.536
K-L	0.879	0.832	0.758	0.840	0.792	0.394	1.354	1.000

Notice that both K-L information and Shannon entropy procedures perform markedly better than random selection in terms of individual attribute and whole attribute pattern classification. Even in the worst case within this simulation study, the K-L information is 32% more accurate than random selection with respect to whole pattern recognition (see Table 3). From Tables 2 to 5, it is clear that the Shannon entropy procedure markedly outperforms the procedure based on K-L information in both aspects. In recognition of the whole pattern the Shannon entropy

Table 3: Correct classification rates for 50-item-5-attribute test (2500 trials)

	1	2	3	4	5	whole pattern	whole pattern ratio to 'random'	whole pattern ratio to 'K-L'
random	0.840	0.803	0.806	0.833	0.809	0.418	1.000	0.757
Shannon	0.949	0.943	0.902	0.942	0.926	0.750	1.794	1.359
K-L	0.928	0.891	0.824	0.901	0.854	0.552	1.320	1.000

Table 4: Correct classification rates for 30-item-8-attribute test (2500 trials)

	1	2	3	4	5	6	7	8	whole pattern	whole pattern ratio to 'rand'	whole pattern ratio to 'K-L'
random	0.764	0.782	0.732	0.772	0.745	0.763	0.754	0.763	0.167	1.000	0.485
Shannon	0.911	0.935	0.910	0.966	0.910	0.959	0.920	0.941	0.653	3.910	1.900
K-L	0.914	0.874	0.732	0.949	0.808	0.926	0.824	0.888	0.344	2.060	1.000

Table 5: Correct classification rates for 50-item-8-attribute test (2500 trials)

	1	2	3	4	5	6	7	8	whole pattern	whole pattern ratio to 'rand'	whole pattern ratio to 'K-L'
random	0.801	0.836	0.801	0.831	0.805	0.832	0.801	0.832	0.243	1.000	0.443
Shannon	0.975	0.981	0.962	0.989	0.962	0.982	0.967	0.976	0.833	3.427	1.520
K-L	0.938	0.932	0.861	0.970	0.867	0.969	0.865	0.940	0.548	2.255	1.000

procedure outperforms the K-L procedure by 36% to 90%(see Table 3 and Table 4). We observed that the 8-attribute item bank results in higher classification rates than under the 5-attribute item bank.

The averages of individual attribute classification rates under different selection rules in four testing contexts are listed in Table 6. It can be seen that the Shannon entropy procedure has higher classification accuracy than KL, which in turn has higher classification accuracy than random selection. Table 7 shows the distributions of ex-

Table 6: Correct classification rates averaged over attributes (2500 trials)

	random	K-L	Shannon
30-item,5-attribute	0.761	0.82	0.887
50-item,5-attribute	0.818	0.880	0.932
30-item,8-attribute	0.759	0.864	0.932
50-item,8-attribute	0.817	0.918	0.974

posure rates of different methods in terms of the mean classification rate and various quantiles of classification rates. For example, in the case of 30 items and 5 attributes, we see that the exposure rate under random selection is uniform, as expected. The exposure rate under K-L selection is right skewed, with the third quantile being much smaller than the maximum. This implies that a large proportion of items in the item bank have a very low chance to be used. In this respect, the Shannon entropy is even worse than K-L selection. By the Shannon entropy procedure, the maximum item exposure probability is 0.932, while the third quartile is only 0.03. This very skewed distribution of item exposure rates under the Shannon entropy procedure and under KL selection is also seen in the remaining three cases.

5 Discussion

Both procedures are promising in that they result in very high correct classification rates for individual attributes and whole attribute patterns. However, they both have very high item exposure rates for certain items. This problem is of critical

Table 7: summary statistics of item exposure rates using Shannon or KL (proportion of exams in which item appears)

		Min	1st Quartile	Median	Mean	3rd Quartile	Max
30-item-5-attribute	Random	0.050	0.060	0.062	0.063	0.066	0.080
	K-L	0.000	0.000	0.007	0.063	0.041	0.823
	Shannon	0.000	0.000	0.000	0.063	0.030	0.932
50-item-5-attribute	Random	0.094	0.109	0.112	0.112	0.116	0.138
	K-L	0.000	0.000	0.031	0.112	0.118	0.919
	Shannon	0.000	0.001	0.018	0.112	0.121	0.943
30-item-8-attribute	Random	0.045	0.056	0.059	0.059	0.062	0.072
	K-L	0.000	0.000	0.001	0.059	0.045	0.921
	Shannon	0.000	0.000	0.001	0.059	0.031	0.973
50-item-8-attribute	Random	0.091	0.103	0.107	0.107	0.111	0.126
	K-L	0.000	0.000	0.026	0.107	0.098	0.968
	Shannon	0.000	0.001	0.020	0.107	0.128	0.981

importance in CAT. For over a decade, many psychometricians have worked to develop methodology to control exposure rates without greatly reducing the efficiency of item selection procedures (Stocking and Lewis, 2000; Chang and Ying, 1996). The fact that the exposure rates under the Shannon entropy procedure can be as high as 0.97, meaning some items were used for as many as 97% of the examinees, and nearly 60% of items in the item bank were left unused in the entire simulation warns us to apply Shannon entropy with great care in real situations. An efficient procedure that controls exposure rates has to be developed before CAT is practical in cognitive diagnosis.

In our study, we also studied two additional methods to select items for CAT. They both basically select the next item by maximizing the weighted posterior of the current attribute pattern estimate. It turned out that these two methods did not perform well, and they are not included in this paper. An important direction of future research is to find a way to use the Shannon entropy procedure and/or K-L information that can result in more balanced item exposure. The current study indicates that the Shannon entropy procedure is superior and modifications that allow a greater distribution of item exposure are desirable.

References

[1] Birnbaum A.(1968). Some latent trait models and their use in inferring an examniee's ability. In F.M.Lord and M.R. Novick, *Statistical theories of mental test*

scores, 397-479. NA: Addison-Wesley.

[2] Chang H.H and Ying Z. (1996). A global information approach to computerized adaptive test. *Applied Psychological Measurement*, 20, 213-229.

[3] Cover, T.M. and Thomas, J.A. (1991). *Elements of information theory*, Wiley-Interscience.

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

[5] DiBello,L.V.,Stout,W.F.,Roussos,L.A.(1995). Unified cognitive / psychometric diagnostic assessment likelihood-based classification techniques. In P.D.Nichols, S.F.Chipman and R.L.Brennan(Eds.), *Cognitively diagnostic assessment*, Lawrence Erlbum Associates, Chapter 15,361-389.

[6] Hartz,S.(2002). A Bayesian framework for the Unified model for asessing cognitive abilities: blending theory with practicality. Doctoral thesis in University of Illinois.

[7] Junker, B. and Sijtsma, K.(2001). Cognitive assessment models with few assumptions and connections with non-parametic item response theory. *Applied Psycholigical Measurement*, 25,258-272.

[8] Lehmann E.L. and Casella, G. (1998). *Theory of point estimation.* New York: Springer Verlag.

[9] Maris,E.(1999). Estimating multiple classification latent class models, *Psychometrika* 64, 187-212.

[10] Roussos,L.(1994). Summary and review of cognitive diagnosis models. Unpublished Manuscript.

[11] Stocking,M.L. and Lewis,C.(2000). Methods of controlling the exposure of items in CAT. In Van der Linden and Glas,C.A.(Eds.),*Computerized adaptive testing: theory and practice*. Kulwer Academic Publishers.

[12] Tatsouka,K.K.(1984). Caution indices based on item response theory. *Psychometrika*,49,95-110.

[13] Tatsouka,C. and Ferguson,T.(2003). Sequential classification on partially ordered set. Manuscript.

[14] Tatsouka,C.(2002). Data analytic methods for latent parially ordered classification models. *Journal of Royal Statistical Society*, 51, 337-350.

Figure 1: the Shannon entropy vs. p for Bernoulli variable

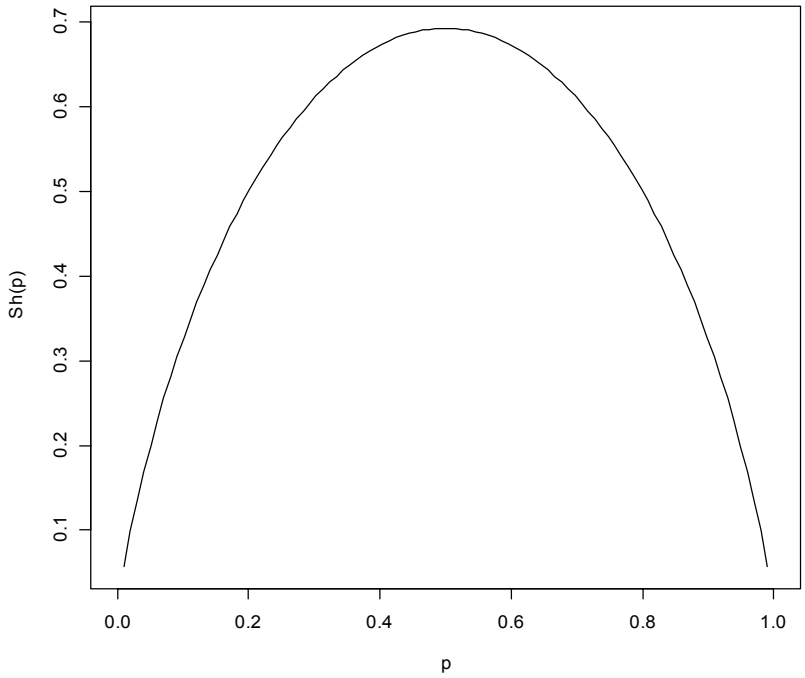

Figure 2: The distribution of attributes measured per item (item bank 1)

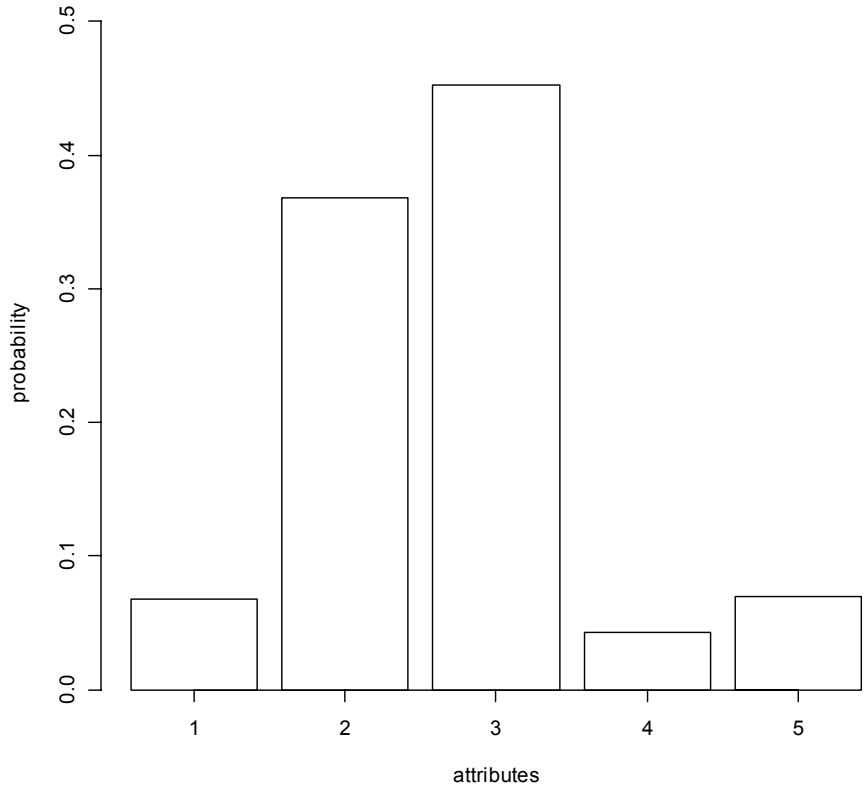

Figure 3: The distribution of attributes measured per item (item bank 2)

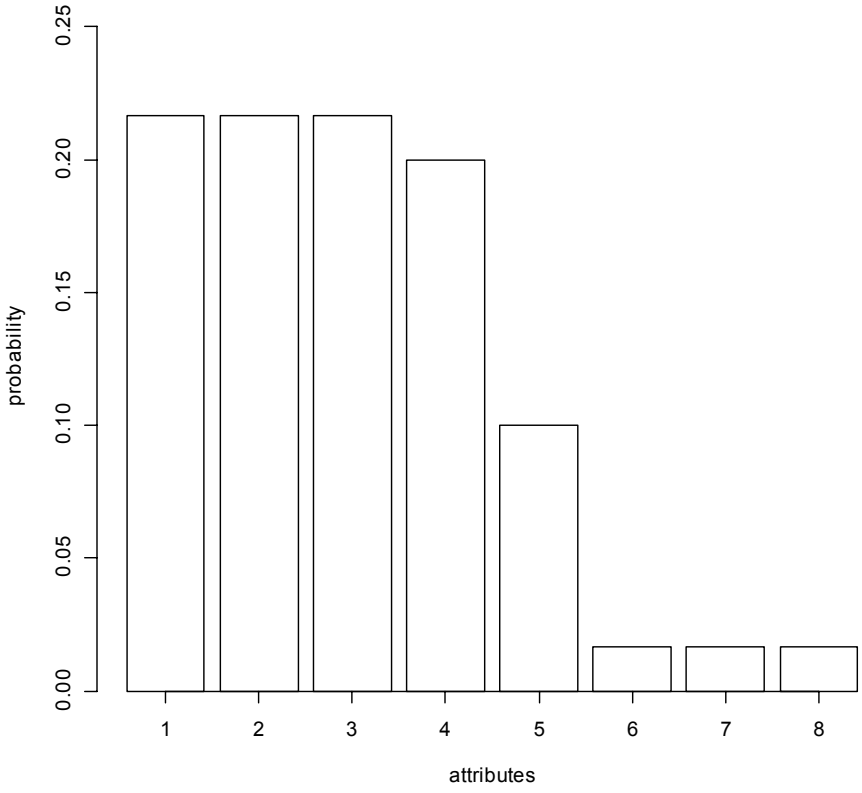

Figure 4 The distribution of r* in item bank 1 for each attribute

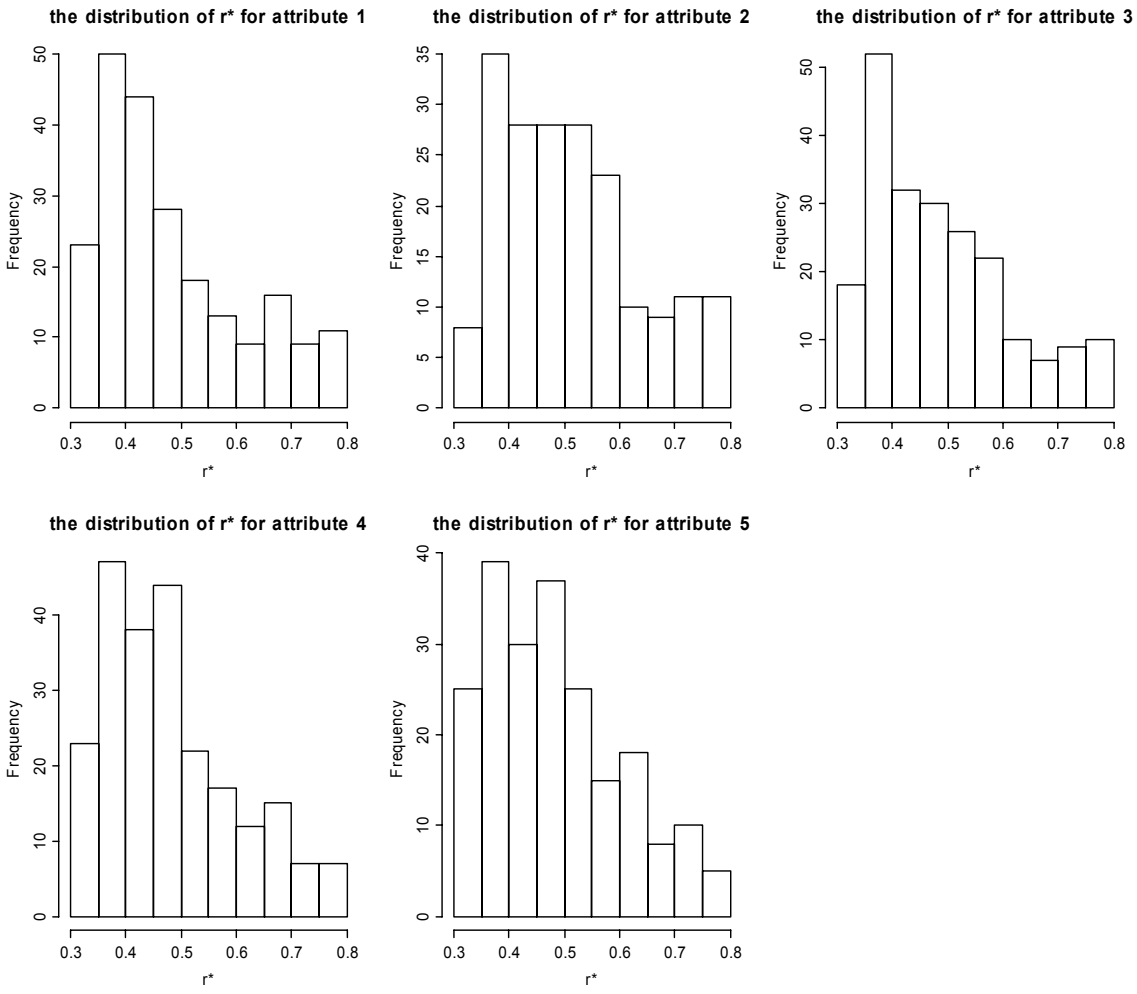

Figure 5 the distribution of r* in item bank 2 for each attribute