The Korean Communications
in Statistics Vol. 8, No. 1, 2001
pp. 193-207

Development of an Item Selection Method for Test-Construction by using a Relationship Structure among Abilities[1)]

Sung Ho Kim[2)], Mi Sook Jeong[3)] and Jung Ran Kim[4)]

Abstract

When designing a test set, we need to consider constraints on items that are deemed important by item developers or test specialists. The constraints are essentially on the components of the test domain or abilities relevant to a given test set. And so if the test domain could be represented in a more refined form, test construction would be made in a more efficient way. We assume that relationships among task abilities are representable by a causal model and that the item response theory (IRT) is not fully available for them. In such a case we can not apply traditional item selection methods that are based on the IRT. In this paper, we use entropy as an uncertainty measure for making inferences on task abilities and developed an optimal item selection algorithm which reduces most the entropy of task abilities when items are selected from an item pool.

Keywords : Ability states; Bayes network; Conditional independence; Conditional probability; Entropy; Item response theory; Test information

1. Introduction

In recent years test construction paradigms have been proposed under IRT framework, where items are selected in such a way that some aspect of the items to be selected is optimized subject to constraints on other aspects. The constraints may incorporate item information function and characteristics of items, such as contents or types, that are deemed important by test developers or subject experts. Literature in this line of work includes van der Linden (1987), van der Linden and Boekkooi-Timminga (1989), Ackerman (1989), Stocking and

1) The authors wish to acknowledge the financial support of the Korea Research Foundation made in the program year of 1998.
2) Associate Professor, Division of Applied Mathematics, KAIST, Daejeon, 305-701, Korea.
E-mail : shkim@sorak.kaist.ac.kr
3) Part-time lecturer, Hannam University, Daejeon, 300-791, Korea.
4) Professor, Statistical Training Center, National Institute of Professional Administration, Daejeon, 305-703.

Swanson (1993), and Swanson and Stocking (1993).

Tests are usually evaluated in two perspectives, reliability and validity. No matter what the purposes of a test are, it is desirable that the two concepts be substantially well tuned to the purposes of the test. While the notion of test reliability is quite technical, test validity has been described from a variety of aspects. It is noteworthy that Cronbach (1980), Dunnette and Borman (1979), Guion (1977, 1978), Messick(1975, 1994), and Tenopyr (1977) among others stated forcefully to the effect that the different types of validity, including content, criterion, and construct validities, are inseparable in essense and that all types of validation are one and, in a sense, are rooted in construct validation.

Embretson (1983, 1985) considered a conceptual model for test design that shows a relationship of the cognitive features of items such as task abilities and strategies to construct validity. The model includes multicomponent latent trait models (MLTMs) which provide estimates of the cognitive demands in each item and specify the relationship of cognitive demands to the cognitive abilities that are reflected in item solving. (p. 195, Embretson, 1985) The relationship among the task abilities and the item scores involved in an MLTM is representable by a graph of nodes (or vertices) and edges, where a node represents a variable and a pair of nodes are connected by an edge if the variables corresponding to the nodes are associated. The relationship among task abilities, whether they are knowledge stores, strategies, or problem-solving abilities, are pre-requisite in general, and the relationship between an item and its relevant abilities is of cause-effect type. Thus, the relationship among abilities and items can be represented via directed acyclic graph (DAG). In educational testing, such a graph was named *Bayesian Inference Network* (BIN) in Mislevy (1994), which is a mixture of *Bayes Network* (Pearl, 1988) and *inference*. Bayes Network is another name of DAG, which is used to represent the cause-effect relation among variables, and "inference" is made on abilities and/or item scores given evidence on a set of variables involved in a given graph or model.

A generalized version of the MLTM is a recursive model (Wermuth and Lauritzen, 1983). The joint probability for a recursive model can be expressed in a factorized form, where each factor is a marginal or conditional probability of a variable involved in the model conditional on the conditioning variables of the variable. The model structure of any recursive model is representable by a DAG. The pre-requisite or its equivalent relationship among task abilities and the cause-effect relationship between a set of items and a set of their item-relevant abilities can be depicted in a BIN. Arrows go from the node of a lower-level ability to the node of a higher-level one, and arrows also go to the node of an item from those of the item-relevant abilities.

If a BIN of test items and item-relevant abilities well fits to a test data set, it is an evidence supporting the construct validity of the test where the construct is substantiated in the BIN. Although the test construct is well validated, that is, the structure of the BIN is well supported by data, the test may yet have to undergo another evaluation in "inference accuracy." Accuracy in making inferences for task abilities has much to do with test reliability whether the latter is defined under classical test theory or item response theory. We aim in this paper to explore item selection rules for an optimal test design. "Optimality" is in the sense that a prediction for ability states of a test-taker by a test set is more accurate than

another test-set.

This paper consists of 6 sections. In section 2, we briefly consider inference making under the IRT frame, and then an uncertainty measure entropy is introduced as a test information measure along with some of its basic properties. Section 3 presents a theoretic result to the effect that one can select, under a certain condition, a most informative item using entropy as an uncertainty measure. In section 4, comparisons are made by simulation among items some of which tap one ability and the others tapping two or three abilities. In the simulation study, we considered a variety of relationships among ability variables. Section 5 presents an illustration of optimal item selection where the relationships among abilities are represented via a causal model, and section 6 concludes the paper.

2. Inference Making and Entropy as a Test Information Measure

One of the major purposes of educational testing is making inferences on an interested set of attributes of test takers. Under the IRT, the precision is represented in terms of $I(\theta)$, the information function, or the conditional error variance of θ, the person parameter, and the precision varies across θ.

In adaptive testing under a framework of IRT, items are selected under a certain condition so that predictions on θ may be made as accurate as possible. The condition is usually provided by test specialists in the form of item or test specifications (Swanson and Stocking, 1993). Whether a test is adaptive or conventional, items are selected to attain the same goal that inferences are made with as small error as possible. If making inferences on θ under an IRT frame was a main goal of a test, then the test should be designed so that the test information be as large as possible. But if we were ambitious to make inferences on something far more refined than θ such as task abilities that are required in solving the items of a test set, we might need some other information measure instead of the item or test information function.

The Shannon entropy (or entropy for short) of X can be interpreted as representing the amount of uncertainty that exists in the value of X. We denote by $H(Y \mid X)$ the entropy of Y conditional on X .

The theorem below follows immediately from the definition of entropy, and see section 4.4 of Whittaker(1990) for its proof.

Theorem 2.1 *Let X and Y denote two distinct random vectors of item scores. Then we have*

$$H(A \mid X) \geq H(A \mid X, Y).$$

This theorem is applied to the situation where random variables X and Y are conditionaly independent given Z, i.e.,

$$X \perp Y \mid Z. \tag{2.1}$$

Corollary 2.1 *Suppose that random variables X, Y, and Z satisfy expression (2.1). Then, it holds that*

$$H(Y \mid X) \geq H(Y \mid Z).$$

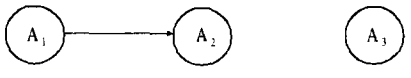
Proof: The proof is immediate from Theorem 2.1 and the condition of the corollary. □

Corollary 2.1 states that in predicting for Y, Z is more informative than X.

3. Entropy and Item Selection

Under IRT, the test information function of a test set is expressed as the sum of the item information functions of the items in the test set. The linear relation between item information and test information makes the test construction quite feasible. But if we intend to predict a more refined version of θ, i.e., the states of the abilities required in a test, we need to consider both the prediction accuracy and the set of abilities that are tapped by each item.

Suppose we want to construct a test set of two items only to make inferences on three task abilities. Denote the states of abilities 1, 2, and 3 by A_1, A_2, and A_3, and suppose that they are related as in the figure below, that is, A_1 and A_2 are associated each other and independent of A_3.



Here our problem is what items to select. There are as many as 7^2 possible test designs, that is, 7 possible ways of ability tapping for each item. To name a few of them, (1) item 1 tapping A_1 and the other tapping the rest; (2) item 1 tapping A_1 and A_2 and the other A_2 and A_3; (3) item 1 tapping A_1 and A_2 and the other A_3. In designs 1 and 2, item 2 taps two abilities that are independent each other while it is not the case for design 3. Is there any criterion by which one can evaluate test designs with respect to accuracy in making inferences? We will explore part of this issue in this section.

By the property of entropy, we have

$$H(X)-H(X \mid A) \geq H(Y)-H(Y \mid A).$$

In practice, $H(X)-H(X \mid A)$ is easier to compute than $H(A)-H(A \mid X)$, because to obtain $H(A \mid X)$ we need to compute the joint conditional probability of A given X while $H(X \mid A)$ is obtained, assuming that the test consists of I items, as

$$H(X \mid A)=\sum_{i=1}^{I} H(X_i \mid A),$$

where X_i is the item score variable of item i. This equation is possible since X_i is conditionally independent of X_j, $j \neq i$, given A. So we will use $H(X)-H(X \mid A)$ instead of $H(A)-H(A \mid X)$ in comparing the uncertainty levels between tests.

We will denote the number of the task abilities that are required for a given test item by K and by A the vector of the states of the K abilities. We will call by a k-tapping item an item that taps k task abilities, that is, an item which is causally related with the k abilities. For notational convenience, We will write P_k and $P_{k|j}(l)$ for $P(A_k=1)$ and

$P(A_k=1 \mid A_j=l)$, respectively, and write $X_{(j\cdots k)}$ for the item score variable of the item that taps abilities $j,\cdots,k$ and $A_{j\cdots k}$ for $(A_j,\cdots,A_k)$. We define $H_{j\cdots k}$ by

$$H_{j\cdots k}=H(X_{(j\cdots k)})-H(X_{(j\cdots k)} \mid A_j,\cdots,A_k) \quad \text{for } 1\leq j\leq K.$$

We will also use $CP_{j\cdots k}(a)$ for $P(X_{(j\cdots k)}=1 \mid (A_j,\cdots,A_k)=a)$.

The following result is immediate from the definition of entropy.

Theorem 3.1 *Suppose that the three random vectors X, Y, and Z satisfy the conditional independence,*

$$X\perp Z \mid Y.$$

Then it follows that

$$H(X,Y)-H(X,Y \mid Z)=H(Y)-H(Y \mid Z).$$

Proof: By the conditional independence, we have

$$H(X,Y \mid Z)=H(Y \mid Z)+H(X \mid Y).$$

So follows the desired result. □

Since the conditional probability of $X_{(j\cdots k)}$ is influenced by $A_{j\cdots k}$ only, we can reexpress $H_{j\cdots k}$, by Theorem 3.1, as

$$H_{j\cdots k}=H(X_{(j\cdots k)})-H(X_{(j\cdots k)} \mid A_{j\cdots k}).$$

The theorem is also applied in deriving a rule to select an item out of a set of 1-tapping items.

Theorem 3.2 *Suppose that $K\geq 2$ and that items 1 and 2 tap abilities 1 and 2, respectively. Suppose also that A_i and X_i, $i=1,2$, are binary, each taking on 0 or 1, and that*

$$P(X_1=1 \mid A_1=1)=P(X_2=1 \mid A_2=1) \tag{3.1}$$

and

$$P(X_1=1 \mid A_1=1)+P(X_1=1 \mid A_1=0)=1. \tag{3.2}$$

Then we have the following:

$$H_1\geq H_2 \quad \textit{if and only if} \quad H(A_1)\geq H(A_2), \tag{3.3}$$

where equality is a simultaneous occurrence on both sides.

Proof: From (3.1) and (3.2) follows that, for $i=1,2$,

$$H(X_i \mid A_i=1)=H(X_i \mid A_i=0).$$

So we have

$$\begin{aligned} H(X_1 \mid A_1)-H(X_2 \mid A_2) &= \sum_{a=0}^{1}(P(A_1=a)-P(A_2=a))H(X_1 \mid A_1=a) \\ &= H(X_1 \mid A_1=1)\sum_{a=0}^{1}(P(A_1=a)-P(A_2=a)) \\ &= 0 \end{aligned} \tag{3.4}$$

This result yields

$$H_1-H_2=H(X_1)-H(X_2). \tag{3.5}$$

Denote $P(X_1=1 \mid A_1=1)$ by α and $P(A_i=1)$ by P_i. Then, from (3.1) and (3.2), we

have, for $i=1,2$,

$$P(X_i=1)=1-a+P_i\cdot(2a-1).$$

$P(X_i=1)$ is linear in P_i and equal to 0.5 when $P_i=0.5$. Thus it follows that

$$|P(X_1=1)-0.5|\geq|P(X_2=1)-0.5|$$

if and only if

$$|P_1-0.5|\geq|P_2-0.5|,$$

where equality holds only simultaneously. This implies the result (3.3) by (3.5), since the function

$$f(x)=x\log x+(1-x)\log(1-x),\quad 0\leq x\leq 1,$$

is convex and symmetric about $x=0.5$.□

According to Theorem 3.2, when all the variables are binary and all the items are 1-tapping, it is desirable to tap the ability node whose marginal entropy is the largest provided that the conditions (3.1) and (3.2) are satisfied for all the items. But the theorem can not be extended to multi-tapping items. For instance, consider a 2-tapping situation. For item 1 which taps a pair of abilities A_1 and A_2, we can imagine an extended version of (3.2) such as

$$P(X_1=1\mid A_1+A_2=2)+P(X_1=1\mid A_1=a_1,A_2=a_2)=1, \tag{3.6}$$

for $0\leq a_1+a_2\leq 1$. Suppose that item 2 taps another pair of abilities, A'_1 and A'_2, that

$$P(X_1=1\mid A_1+A_2=2)=P(X_2=1\mid A'_1+A'_2=2),$$

and that an analogy of (3.6) holds for X_2. Then we can easily see that

$$H(X_1\mid A_1,A_2)-H(X_2\mid A'_1,A'_2)=0$$

which is an analogy to (3.4). But it is not guaranteed that

$$(H(X_1)-H(X_2))(H(A_1,A_2)-H(A'_1,A'_2))\geq 0,$$

because $H(A_1,A_2)$ ranges over a non-empty interval for the set of $P(A_1,A_2)$ values that satisfy

$$\begin{aligned}P(X_1=1) \;=\; & (1-P(A_1=1,A_2=1))(1-P(X_1=1\mid A_1=1,A_2=1))\\ & +P(A_1=1,A_2=1)P(X_1=1\mid A_1=1,A_2=1).\end{aligned}$$

Therefore, Theorem 3.2 can not be extended to multi-tapping situations. However, if the marginals of X are available, there is yet a hope.

Theorem 3.3 *Consider a test set which requires at least t task abilities and suppose that items 1 and 2 are t-tapping and that item i ($i=1,2$) taps the set $A^{(i)}$ of ability variables. Suppose also that X's and A's are all binary, each taking on 0 or 1, and that*

$$P(X_1=1\mid A^{(1)}\ \textit{is a vector of}\ 1\text{'s})=P(X_2=1\mid A^{(2)}\ \textit{is a vector of}\ 1\text{'s}) \tag{3.7}$$

and

$$P(X_1=1\mid A^{(1)}\ \textit{is a vector of}\ 1\text{'s})+P(X_1=1\mid A^{(1)}=a)=1, \tag{3.8}$$

for every t-vector a whose components are not all 1. Then we have the following:

$$H_1 - H_2 = H(X_1) - H(X_2). \tag{3.9}$$

Proof: The proof is a straightforward application of the argument that leads to (3.4). □

If an item is solvable when and only when the states of all the abilities that are tapped by the item are good enough for the item, we will call the item *all-or-fail* item. If all the item-relevant abilities are in good states, we will say that they are in a *perfect* state; otherwise, in an *imperfect* state. The item whose probabilities of correct response are the same across the imperfect states of the item-relevant abilities is of all-or-fail type.

The item which satisfies (3.2) or (3.8) is not unusual in the real world. For instance, if it is 2-tapping, we can think of the case that $CP_{12}(11)=0.85$ and $CP_{12}(00)=CP_{12}(01)=CP_{12}(10)=0.15$. When it comes to 1-tapping, the conditions of Theorem 3.2 may often be satisfied, in which case we can apply this theorem in selecting a most informative item provided that the uncertainty level for each ability is available. If the marginals of item scores are available, then Theorem 3.3 is useful for item-selection.

If items are not of all-or-fail type, preference between items is subject to the joint probability distribution of A and the conditional distribution of X given A. We will explore this for some situations where items tap as many as 3 abilities.

4. Item Selection In 3 Simple Situations where $K \leq 3$

In this section, we will consider the following three situations:

Situation 1: $K=2$, and 1-tapping and 2-tapping will be compared.

Situation 2: $K=3$ and $A_1 \perp A_3 \mid A_2$. Cmparisons will be made on the 2-tappings on the A_1-A_2 pair and on the A_1-A_3 pair.

Situation 3: $K=3$ and $(P_{3|12}(0,0), P_{3|12}(0,1), P_{3|12}(1,0), P_{3|12}(1,1))=(0.1,0.3,0.3,0.9)$. The preferences among the three 1-tappings, three 2-tappings, and a 3-tapping will be ranked.

Situation 1

Preference between 1-tapping item and 2-tapping item is subject to the probability distribution of A and the conditional distribution of X given A as illustrated in Table 4.1. The table is obtained under the condition that $(CP(0), CP(1))=(0.15, 0.85)$ for a 1-tapping item and $(CP(0,0), CP(0,1), CP(1,0), CP(1,1))=(0.1, 0.3, 0.3, 0.9)$ for a 2-tapping item. According to the table, the item that taps A_1 was most preferable under Model 1, as for Model 2 it was the item that taps A_2, and it was the item that taps both A_1 and A_2 as for Model 3.

To get an insight into the preferences between 1-tapping and 2-tapping, we obtained rank-averages between the tappings with the same conditional probabilities of X as for Table 4.1. Table 4.2 shows the rank-averages for each of $P(A_1=1)=0.1(0.1)0.9$, where the averages were taken over the 45 joint probabilities of A_1 and A_2, $P_{2|1}(0)=0.1(0.1)0.9$ and $P_{2|1}(1)=P_{2|1}(0)(0.1)0.9$. According to the table, 1-tapping was most preferable on

average when $P_1 = 0.1(0.1)0.6, 0.9$. When $P_1 = 0.8$, H_2 was very close to H_{12}. A possible

Table 4.1: Values of $H(X) - H(X \mid A)$ in situation 1. $(CP(0), CP(1)) = (0.15, 0.85)$ and $(CP_{12}(0,0), CP_{12}(0,1), CP_{12}(1,0), CP_{12}(1,1)) = (0.1, 0.3, 0.3, 0.9)$.

Model	P_1	$P_{2\mid 1}(0)$	$P_{2\mid 1}(1)$	H_1	H_2	H_{12}
1	0.70	0.20	0.20	0.231	0.179	0.136
2	0.70	0.30	0.40	0.231	0.254	0.211
3	0.70	0.30	0.90	0.231	0.222	0.285

explanation of this is that when P_1 is as small as 0.1 or 0.2 or as large as 0.8 or 0.9, tapping A_2 only may be most informative, since the uncertainty level for the state of A_1 is relatively low. In the same context, we could understand that H_1 was the largest when $P_1 = 0.4, 0.5, 0.6$.

The ranking among H_1, H_2, and H_{12} is subject to the joint distribution of A_1 and A_2 and the conditional probability distribution of X given A. We will see below what happens when A_1 and A_2 become more associated each other. For notational convenience, let

$$H_{A_1,A_2}(X \mid a_1, a_2) = H(X \mid A_1 = a_1, A_2 = a_2).$$

Assuming that A_1 and A_2 are binary, we have

$$H(X \mid A_1, A_2) = \sum_{a_1, a_2 = 0}^{1} H_{A_1,A_2}(X \mid a_1, a_2) P_1(a_1) P_{2\mid 1}(a_1)^{a_2} (1 - P_{2\mid 1}(a_1))^{1-a_2}.$$

Hence

$$\frac{\partial H(X \mid A_1, A_2)}{\partial P_{2\mid 1}(0)} = P_1(0)(H_{A_1,A_2}(X \mid 0,1) - H_{A_1,A_2}(X \mid 0,0))$$

and

$$\frac{\partial H(X \mid A_1, A_2)}{\partial P_{2\mid 1}(1)} = P_1(1)(H_{A_1,A_2}(X \mid 1,1) - H_{A_1,A_2}(X \mid 1,0)).$$

Provided that

$$H_{A_1,A_2}(X \mid 0,0) < H_{A_1,A_2}(X \mid 0,1) \text{ and } H_{A_1,A_2}(X \mid 1,0) > H_{A_1,A_2}(X \mid 1,1), \quad (4.1)$$

we have that

$$\frac{\partial H(X \mid A_1, A_2)}{\partial P_{2\mid 1}(0)} > 0 \text{ and } \frac{\partial H(X \mid A_1, A_2)}{\partial P_{2\mid 1}(1)} < 0.$$

That is, for a given $H(X_{(1\,2)})$, H_{12} increases under condition (4.1) as A_1 and A_2 become more associated each other. Condition (4.1) seems very likely in real world. It is not unusual that $CP(0,0) < 0.5 < CP(1,1)$ and that $\min\{\mid CP(0,0) - 0.5 \mid, \mid CP(1,1) - 0.5 \mid\} > \max\{\mid CP(a_1, a_2) - 0.5 \mid ; a_1 + a_2 = 1\}$ which lead to condition (4.1).

Situation 2

In this situation, We considered only 2-tapping items to compare the uncertainty level,

$H(X)-H(X \mid A)$, between tapping both of A_1 and A_2 and tapping both of A_1 and A_3, but the preference is subject to the joint distribution of A and the conditional distribution of X given A as illustrated in Table 4.3.

Table 4.2: Rank-averages of H_1, H_2 and H_{12} for each of $P_1=0.1(0.1)0.9$.

P_1	H_1	H_2	H_{12}
0.10	1.689	2.956	1.333
0.20	1.933	2.733	1.311
0.30	20111	2.444	1.422
0.40	2.356	1.933	1.644
0.50	2.622	1.467	1.822
0.60	2.222	1.711	2.000
0.70	1.689	2.022	2.267
0.80	1.244	2.333	2.378
0.90	1.022	2.600	2.356

Table 4.3: Values of $H(X)-H(X \mid A)$ in situation 2. For $j=2,3$, $(CP_{1j}(0,0), CP_{1j}(0,1), CP_{1j}(1,0), CP_{1j}(1,1))=(0.1,0.2,0.2,0.9)$.

Model	P_1	$P_{2\mid1}(0)$	$P_{2\mid1}(1)$	$P_{3\mid2}(0)$	$P_{3\mid2}(1)$	H_{12}	H_{13}
1	0.7	0.35	0.65	0.15	0.55	0.31	0.25
2	0.7	0.35	0.95	0.05	0.75	0.30	0.32

The proportions of the cases that tapping both of A_1 and A_2 is preferable to tapping A_1 and A_3 for a set of distributions of A are as follows, where it is assumed that $CP_{12}=CP_{23}$. As for
the joint distributions of A, we considered $P_1=0.1(0.1)0.9$, $P_{2\mid1}(0)=0.05(0.1)0.45$, $P_{2\mid1}(1)=0.55(0.1)0.95$, $P_{3\mid2}(0)=0.05(0.1)0.45$, and $P_{3\mid2}(1)=0.55(0.1)0.95$. What we found is that the preference of tapping the A_1-A_2 pair is around or above the proportion 0.9 except the four distributions of A; one when $P_1=0.1$, another when $P_1=0.8$, and the rest two when $P_1=0.9$.

Situation 3

In this situation, it is assumed that A_3 is causally influenced by both A_1 and A_2, while A_1 and A_2 may or may not be associated each other. Thus, in terms of abilities, A_1 and A_2 are assumed to be prerequisite to A_3. In light of the above two situations, we may anticipate that as the three abilities are more associated each other, 3-tapping may be more

preferable. But the preference is, as in the above situations, subject to the joint distribution of A and the conditional distribution of X given A as indicated in Table 4.4.

Table 4.4: Rank averages of $H(X)-H(X \mid A)$ among the 7 items, $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, $X_{(12)}$, $X_{(23)}$, $X_{(13)}$, and $X_{(123)}$. The conditional probabilities for the items are assigned in four different settings:

Setting 1. $CP_1 = CP_2 = CP_3$, $CP_1(0) = 0.15$, $CP_1(1) = 0.85$; $CP_{12} = CP_{23} = CP_{13}$, $(CP_{12}(0,0), CP_{12}(0,1), CP_{12}(1,0), CP_{12}(1,1)) = (0.1, 0.3, 0.3, 0.9)$; $(CP_{123}(0,0,0), CP_{123}(0,0,1), CP_{123}(0,1,0), CP_{123}(0,1,1), CP_{123}(1,0,0), CP_{123}(1,0,1), CP_{123}(1,1,0), CP_{123}(1,1,1)) = (0.1, 0.2, 0.2, 0.4, 0.2, 0.4, 0.4, 0.95)$.

Setting 2. Same as Setting 1 except that $CP_1(0) = 0.1$ and $CP_1(1) = 0.9$.

Setting 3. Same as Setting 1 except that $CP_{123}(1,1,1) = 0.9$.

Setting 4. Same as Setting 2 except that $CP_{123}(1,1,1) = 0.9$.

Setting	H_1	H_2	H_3	H_{12}	H_{23}	H_{13}	H_{123}
1	2.88	3.60	4.41	2.10	6.05	4.38	4.52
2	4.59	5.38	6.35	1.26	4.29	2.88	3.20
3	3.08	3.85	4.69	2.61	6.26	4.84	2.62
4	4.68	5.51	6.37	1.77	4.50	3.33	1.80

Table 4.4 is obtained under the joint distribution of A and X as specified below:

$$P_1 = 0.1(0.1)0.9,\ P_{2|1}(0) = 0.1(0.1)0.9,\ P_{2|1}(1) = P_{2|1}(0)(0.1)0.9 \tag{4.2}$$

and $P_{3|12}$ is fixed as

$$(P_{3|12}(0,0), P_{3|12}(0,1), P_{3|12}(1,0), P_{3|12}(1,1)) = (0.1, 0.3, 0.3, 0.9). \tag{4.3}$$

So 405 different joint distributions are considered.

We compared $H(X)-H(X \mid A)$ among the three 1-tapping items, the three 2-tapping items, and one 3-tapping item, where the conditional probabilities are assigned in 4 different settings (see Table 4.4). The values of $H(X)-H(X \mid A)$ are ranked in ascending order for each joint distribution of A and the ranks are averaged over the 405 joint distributions of A for each item.

According to Table 4.4, tapping both A_2 and A_3 appeared to be most preferable on average under settings 1 and 3 and under settings 2 and 4 tapping only A_3 was most preferable on average. It is worth noting that the rank of H_{123} dropped, when the conditional probability $CP_{123}(1,1,1)$ was lowered from 0.95 to 0.9, by the amount of 1.9 between settings 1 and 3 and by the amount of 1.4 between settings 2 and 4. The small change (0.05) in the conditional probability made the 3-tapping far less informative.

It is interesting to note that among the ranks of the 1-tapping items, the rank of H_3 was the highest on average and that among the ranks of the 2-tapping items, the rank of H_{23} was

the highest on average. The conditional probabilities for the 1-tapping items satisfy (3.2). Therefore, the result for the three 1-tapping items implies, according to Theorem 3.2, that A_3 is most uncertain on average. Also note that when $(CP_1(0), CP_1(1))$ is switched from (0.15, 0.85) to (0.1, 0.9) the ranks of the 1-tapping items were all higher than the 2-tapping or 3-tapping items.

As for the 2-tapping items, the conditional probabilities do not satisfy (3.8). There is no clear-cut, theory-based explanation for this. We can, however, say that, under the set-up for Table 4.4, tapping both of A_2 and A_3 is most informative among the 2-tapping items. $H(X)-H(X|A)$ may be regarded as a function of the joint distribution of A and the conditional distribution of X given A. Since the joint distribution of A is of 7 cell probabilities, $H(X)-H(X|A)$ is a function of at least 7 factors. After a simple algebra, we can see that any analysis of the function with respect to the factors may lead us to nowhere

Table 4.5: Rank averages for each of $P_1=0.1(0.1)0.9$.

(a) Setting 1

P_1	H_1	H_2	H_3	H_{12}	H_{23}	H_{13}	H_{123}
0.1	2.600	6.400	5.778	1.533	5.622	3.311	2.733
0.2	3.289	5.667	5.489	1.422	5.778	3.489	2.844
0.3	3.822	4.489	4.689	1.644	5.867	4.067	3.400
0.4	4.156	3.222	3.711	1.978	5.867	4.867	4.178
0.5	4.533	2.111	2.978	2.222	6.044	5.156	4.733
0.6	3.111	2.089	3.733	2.422	6.156	5.178	5.267
0.7	1.978	2.244	4.111	2.667	6.289	5.022	5.644
0.8	1.356	2.756	4.378	2.644	6.378	4.556	5.889
0.9	1.044	3.444	4.844	2.400	6.489	3.800	5.956

(b) Setting 2

P_1	H_1	H_2	H_3	H_{12}	H_{23}	H_{13}	H_{123}
0.1	4.267	6.600	6.311	1.200	4.778	2.622	2.200
0.2	4.844	6.267	6.267	1.089	4.400	2.822	2.289
0.3	5.511	5.867	6.089	1.111	4.044	2.933	2.400
0.4	6.222	5.311	5.933	1.156	3.778	2.911	2.622
0.5	6.822	4.600	5.667	1.156	3.778	2.867	2.911
0.6	5.778	4.622	6.556	1.133	3.822	2.733	3.333
0.7	4.400	4.689	6.844	1.133	4.378	2.533	4.000
0.8	2.267	5.000	6.800	1.422	4.733	3.222	4.533
0.9	1.156	5.467	6.689	1.956	4.911	3.267	4.533

unless the conditions of Theorem 3.3 are satisfied. It sounds quite reasonable to expect that preference of 3-tapping may increase as the 3 ability variables become more highly associated among themselves, which was observed under the settings for Table 4.4. But it is not always the case. We can see a counter-example to this expectation when

$$(P_{3|12}(0,0), P_{3|12}(0,1), P_{3|12}(1,0), P_{3|12}(1,1)) = (0.1, 0.2, 0.1, 0.6)$$

and when the CP_{123} under Setting 1 of Table 4.4 is replaced by $(0.1, 0.2, 0.3, 0.4, 0.2, 0.4, 0.6, 0.95)$. We will thus look into a more detailed picture of Table 4.4.

Table 4.5 lists the rank-averages for each of $P_1 = 0.1(0.1)0.9$. Under setting 1 (see panel a), the 1-tapping of ability 2 was most preferable on average when $P_1 = 0.1$; on the other hand, the 2-tapping of abilities 2 and 3 was most preferable on average when $P_1 = 0.2(0.1)0.9$. But when CP_1 switched from setting 1 to setting 2, the 1-tapping items only were preferable on average. When $P_1 = 0.1, 0.2, 0.3$, tapping ability 2 or 3 was most preferable, and when $P_1 = 0.6(0.1)0.9$, tapping ability 3 was most preferable on average. Tapping ability 1 was most preferable on average when $P_1 = 0.4, 0.5$. The rank-averages among the 1-tapping items are a reflection of Theorem 3.2.

We have considered BINs where abilities are connected by directed edges. The edges are directed not to reflect the cognitive sequence of problem solving activity but to reflect the causal or prerequisite relationship among the variables. In this paper, the items are limited as tapping up to 3 abilities, and the above results are recapitulated as follows:

(i) If three abilities A_1, A_2, and A_3 are related so that $A_1 \perp A_3 \mid A_2$, then items that tap the pair of A_1 and A_3 are preferred the least in general among the 2-tapping items for the three abilities.

(ii) The item preference (or $H(X) - H(X \mid A)$) increases as the probability $P(X = 1 \mid$ all the A's are equal to 1) increases or the probability $P(X = 1 \mid$ all the A's are equal to 0) decreases.

(iii) As a pair of abilities become more positively associated, the preference for the item that taps the ability pair increases; the same is expected to hold for 3-tapping items.

(iv) Preference among 1-tapping items is, under the condition of Theorem 3.2, determined by the amounts of uncertainty of the ability variables.

(v) In situation 3, where 1-, 2-, and 3-tappings are compared, the 1-tapping was most preferred when

$$(CP_1(0), CP_1(1)) = (0.1, 0.9) \tag{4.4}$$

provided that the other conditional probabilities are as in setting 1 of Table 4.4.

Statement (v) suggests that when 1-tapping items are well developed and satisfy (4.4), we may select them based on the amounts of uncertainty of ability variables. That we select the item which taps the ability whose uncertainty is the highest is quite natural from the perspective of decision theory where the aim is to derive a decision scheme to minimize the amount of uncertainty of a given predicted object.

5. An Illustration of Optimal Item Selection

In this section we will demonstrate an optimal item selection for an artificial test set where 14 task abilities are required and which consists of 10 items. Suppose that an item pool is

ready for the test set and that the prerequisite relationships among the 14 abilities are given as in Figure 5.1. The ability states are assumed as binary, 0 for a poor state and 1 for a good or perfect state. The conditional probabilities of an ability or an item are assigned in the same manner as in section 4. The item pool contains 183 items, where items tap 1 through 3 abilities in various forms. Some items tap abilities that are closely related and some others tap abilities that are slightly related or not.

The selection result is summarized Table 5.1. We can see in the table that the results (i) through (v) as listed near the end of section 4 are reflected pretty much. According to the table, an item tapping ability 8 is selected initially as most informative, and ability 7 was tapped next. We can see that the abilities are tapped from near the center toward the boundary of the graph in Figure 5.1.

The selected items would vary across item pools and the relationships among the abilities. So there is no telling which ability or abilities be examined first by an item or which next. However, a general trend is that abilities near the middle of a graph are the first selection

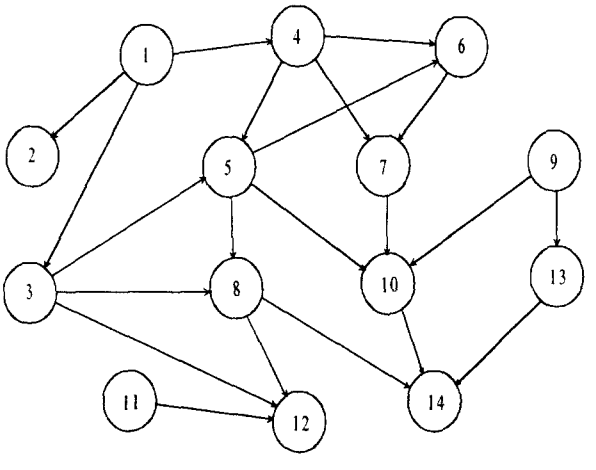


Figure 5.1 A causal model of 14 abilities for an artificial test that is used in section 5.

Table 5.1 List of items that are selected optimally for the 14 abilities as in Figure 5.1

selection sequence	item no	abilities			uncertainty reduction
1	8	8			0.3627D+00
2	7	7			0.2993D+00
3	66	1	2	4	0.2837D+00
4	13	13			0.2630D+00
5	12	12			0.2452D+00
6	10	10			0.2254D+00
7	3	3			0.2172D+00
8	11	11			0.2054D+00
9	41	5	9		0.1880D+00
10	2	2			0.1800D+00

under normal circumstances that the difficulty levels of those abilities are around the medium. The last column in the table shows the reduced amount of uncertainty for making inferences on ability states based on the selected items. The amount is decreasing in the table but it is not necessarily the case in general.

6. Concluding Remarks

An ultimate goal of test design is to produce a test set by which we may predict the states of the test-relevant abilities with the least possible error, i.e., with the smallest possible uncertainty remaining. In this paper, we have explored some fundamentals towards the goal with regard to a fine structure of abilities by considering the graphical structure of abilities and item scores. A more general algorithm for optimal item selection from an item pool can be a straightforward extension of the results of this paper, where the extension may cover various types of items including open ended problems.

It is meaningless comparing a test set that is obtained under the IRT frame with a test set that is obtained under a causal model frame, since the IRT frame is a particular form of the causal model frame. When the relationships among a set of task abilities are represented via a causal model, the item information function under the IRT frame may lead us nowhere.

The computer program for item selection is written in Fortran and it serves well for the multiple choice items.

References

[1] Ackerman, T. (1989). An alternative methodology for creating parallel test forms using the IRT information function. Paper presented at the 1989 NCME annual meeting, San Francisco.

[2] Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement. Measuring Achievement: Progress Over A Decade.* (Ed: William B. Schrader). San Francisco: Jossey-Bass Inc., Publishers. 99-108.

[3] Embretson, S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin*, 93, 1, 179-197.

[4] Embretson, S. (1985). Multicomponent latent trait models for test design. *Test Design: Developments in Psychology and Psychometrics.* (Ed: Susan E. Embretson). Academic Press, Inc. 195-218.

[5] Dunnette, M. C. and Borman, W. C. (1979). Personnel selection and classification systems. *Annual Review of Psychology*, 30, 477-525.

[6] Guion, R. M. (1977). Content validity, the source of my discontent. *Applied Psychological Measurement*, 1, 1-10.

[7] Guion, R. M. (1978). Content validity in moderation. *Personnel Psychology*, 31, 205-214.

[8] Messick, S. A. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-966.

[9] Messick, S. A. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment*, 10, 1, 1-9.

[10] Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 4, 439-483.

[11] Pearl, J. (1988). *Probabilistic Reasoning in Intelligence Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

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

[13] Swanson, L. and Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, 17, 151-166.

[14] Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology*, 30, 47-54.

[15] Wermuth, N. and Lauritzen, S. L.(1983). Graphical and recursive models for contingency tables. *Biometrika*, 70, 3, 537-552

[16] Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, New York: Wiley.