The average Kullback-Leibler (K-L) information index (H. Chang and Z. Ying, in press) is a newly proposed statistic in Computerized Adaptive Testing (CAT) item selection based on the global information function. The objectives of this study were to improve understanding of the K-L index with various parameters and to compare the performance of the K-L index with the traditional information method in CAT item selection. The results of this study, based on simulated and real data with 500 items each, provide evidence that Chang and Ying's global information method produced similar or better true ability theta estimates than the more traditional information approach in CAT item selection. In addition, results from the real item pool analyses indicate the parameter that provides the best theta estimates among the four K-L indices studied. (Contains one table, seven figures, and six references.)

(Author/SLD)
A Comparison of the Traditional Maximum Information Method and the Global Information Method in CAT Item Selection

K. Linda Tang
Educational Testing Service
Princeton, NJ

Paper presented at the annual meeting of the NCME
New York, NY
April, 1996
Abstract

The average Kullback-Leibler (K-L) information index (Chang & Ying, in press) is a newly proposed statistic in CAT item selection based on the global information function. The objectives of this study are: (1) to better understand the performance of the K-L index with different $\delta_n$ values; and (2) to compare the performance of the K-L index with the traditional information method in CAT item selection. The results of this study, based on both simulated and real data, provide evidence that Chang and Ying’s global information method produces similar or better $\theta$ estimates than the more traditional information approach in CAT item selection. In addition, results from the real item pool analyses indicate that the K-L index having $\delta_n = 3/\sqrt{n}$ produces the best $\theta$ estimates among the four K-L indices studied.
A Comparison of the Traditional Maximum Information Method and the Global Information Method in CAT Item Selection

Computerized adaptive testing (CAT) has become popular because this method of testing can provide the same level of measurement precision as conventional paper and pencil testing with the administration of fewer items. This test length reduction advantage provided by CAT is achieved by administering a tailored test to each examinee (Lord, 1971, 1980). CAT selects items that best match an examinee’s ability level. The most popular item selection method used in CAT is the maximum item information method: The (n+1)th item selected for an examinee is the one which provides the maximum information at the examinee’s estimated ability (θn) based on the n items previously administered to that examinee. The item information function is defined in Lord (1980) and is a measure of information in the neighborhood of the θ of interest. Therefore, it is referred to as local information. When the estimated ability θn is not close to the examinee’s true ability (θo), the item which has maximum local information at θn may not be the most appropriate item to administer to the examinee having true ability θo. This will often occur at the early stages of the CAT.

Chang (1995) and Chang and Ying (in press) proposed the use of an alternative item selection method, which they call the global information method, to improve item selection at the early stages of a CAT. The global information function was derived by applying the Kullback-Leibler (K-L) information function in the IRT context. The Kullback-Leibler item information function is defined as the expected value of the likelihood ratio, or the likelihood function at the true ability (θo) divided by the likelihood function at any other ability level θ for an item. It has been shown that the
likelihood ratio statistic is the most powerful statistic to distinguish $\theta_0$ from any other $\theta$ value (Mood, Graybill, & Boes, 1985).

Chang and Ying established the connection between the local and global information functions: The second derivative of the global item information function is the traditional item information function. Geometrically, the traditional information at $\theta = \theta_0$ is the curvature of the global information function at $\theta = \theta_0$. Based on the relationship between the global and local information functions, Chang and Ying developed an information index, the average K-L information index, for CAT item selection. The average K-L information index can be defined as:

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

where $\theta_n$ is the maximum likelihood estimator of $\theta_0$ based on $n$ items administered to an examinee, and $K_j(\theta | \hat{\theta}_n)$ is defined as

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

The parameter $\delta_n$ in (1) generates a sequence which converges to 0 as $n$ increases. The $\delta_n$ controls the width of the interval under the K-L information index curve. This interval is expected to contain the true ability parameter, $\theta_0$, and will narrow down to the neighborhood of $\hat{\theta}_n$ for an examinee as the number of items administered to the examinee increases and as $\delta_n$ approaches $\theta_0$. As Chang and Ying pointed out, for a small $\delta_n$ value, the maximum area
under the K-L curve is equivalent to the maximum curvature, which is the maximum value of the traditional (local) information. For a large $\delta_n$ value, the area is very much influenced by the global information. Therefore, $\delta_n$ is an important parameter in the average K-L index and warrants study under different conditions.

Chang and Ying conducted simulation studies to compare the performance of the local and global information functions in CAT item selection. They concluded that the global information index outperformed the traditional maximum information approach in most of the conditions in their simulation studies. In their study, the $\delta_n$ values used in (1) were $2/\sqrt{n}$ and $1/n$. However, the effect of these two $\delta_n$ values on item selection were not directly compared because these values were used separately in two different simulation studies by Chang and Ying. Also, the test length used in Chang and Ying's study was fixed at 40. The performance of the average K-L index on shorter tests has not been investigated. In addition, in Chang and Ying's study, the simulated item parameters were uniformly distributed, which is often not the case in practice. Therefore, more extensive simulation work need to be carried out to compare the global information method with the traditional information method.

**Objectives**

Because the average K-L information index is a newly proposed statistic, the behavior of this statistic has not been studied extensively using either simulation techniques or empirical data. The purposes of this study were:

1. To better understand the performance of the average K-L index with different $\delta_n$ values; and
2. To compare the performance of the global
information method with the conventional information method in CAT item selection using both real and simulated data.

Methods

1. Data

Both real and simulated data were used in the study. Five hundred Structure and Written Expression items from the Test of English as a Foreign Language (TOEFL) were used to form an item pool. In addition, five hundred simulated items were included in a simulated item pool. The three parameter logistic (3-PL) model was used in both the real and simulated conditions. For the simulated items, the values of the item discrimination parameter (\(a\)) in the 3-PL model were generated from a lognormal distribution (\(\text{LOG}(a) \sim N(0,0.5^2)\)). The values of the item difficulty parameter (\(b\)) in the 3-PL were generated from a \(N(0,2^2)\) distribution. Finally, the values of the \(c\) parameter in the 3-PL were generated from a beta distribution with \(\alpha = 4\) and \(\beta = 13\) (The mean of the beta distribution = 0.24 and the distribution has a weight equivalent to 15 observations of the responses of examinees of very low ability). The distributions of the item discrimination and item difficulty parameters used in this study are also used as the default prior distributions in the computer program BILOG (Mislevy & Bock, 1990) and represent common data structures in practice. In order to make the simulated data to be more representative of the real data, items having \(b\) parameters greater than 3.0 and less than -3.0 or items having a \(a\) parameters greater than 2.5 were eliminated. The beta distribution used to generate the pseudo-guessing or \(c\) parameters provides values similar to those observed in large scale testing programs such as TOEFL. The summary statistics for the item parameters in the real and the simulated item pool are presented in Table 1.
2. True abilities

The following six true abilities ($\theta$s) were used in the study:
-3.0, -2.0, -1.0, -1.0, 2.0, 3.0. These $\theta$ values cover the ability range typically observed in practice.

3. Test length

The test lengths used were 20 for the real data case and 30 for the simulated data case.

4. $\delta_n$ used in the average K-L information index

The following four $\delta_n$ values were used in the real data case: $3/\sqrt{n}$, $1/n$, $1/e^{0.1n}$, and $3/e^{0.1n}$.

The convergence rate (to zero) of these four $\delta_n$ values as a function of $n$ are presented in Figure 1.

Figure 1 illustrates that the four $\delta_n$ sequences converge to 0 at different rates. As mentioned earlier, Chang and Ying pointed out that maximum curvature under K-L curve is the maximum value of the traditional information when the $\delta_n$ value is small. The shifting of K-L to the traditional information occurs when the $\delta_n$ sequence approaches to 0. In other words, the K-L indices associated with each of the four $\delta_n$ sequences shift from global information to local information (traditional information) at
different rates. For example, the K-L index having $\delta_n = 1/n$ shifts from
global to local information around the administration of the 15th item; the K-
L index having $\delta_n = 1/e^{0.1n}$ shifts from global to local information around the
administration of the 25th item; and the K-L index having $\delta_n = 3/e^{0.1n}$ shifts
from global to local information around the administration of the 35th item.
The K-L index having $\delta_n = 3/\sqrt{n}$ clearly converges to 0 much slowly than the
other three $\delta_n$ values.

In summary, there were 30 experimental conditions in the real data case:
6 true $\theta$ values (-3.0, -2.0, -1.0, 1.0, 2.0, and 3.0) x 4 $\delta_n$ values (3/$\sqrt{n}$,
1/n, and 1/e^{0.1n}, and 3/e^{0.1n}) for the global information method plus the 6 true
$\theta$ values for the conventional information method. Because results from the
real data pool analyses indicate that the K-L index having $\delta_n = 3/\sqrt{n}$ produced
the best $\theta$ estimates among the four K-L indices, only the K-L index having $\delta_n$
= 3/\sqrt{n} were used to select items from the simulated item pool. There were 12
conditions in the simulated data case: the 6 true $\theta$ values for both the K-L
index and the conventional information method. The test length for the real
item pool was 20 and for the simulated item pool, 30. One hundred
replications were conducted at each of the six ability levels for each of
these experimental conditions.

5. Analyses

The bias and mean squared error at each of the six ability levels for
each of the experimental conditions were compared.

The bias for the $\theta$ estimate after item i is administered to examinee j
having true ability $\theta_0$ is defined as
and mean squared error for the $\theta$ estimate after item $i$ is administered to examinee $j$ having true ability $\theta_0$ is defined as

$$MSE_i = \frac{1}{100} \sum_{j=1}^{100} (\hat{\theta}_{ij} - \theta_0)^2,$$  

where $\hat{\theta}_{ij}$ is the estimate for $\theta_0$ obtained using either the global information method or the conventional information method, and $i = 3, \ldots, 20$ or 30. The Bias and MSE for the first two items were not computed because these items were not selected by the information methods.

Results

Real Data

Figures 2 and 3 present the bias and MSE at the six true $\theta$ values for the K-L index and the conventional information index (INFO). Each of the four $\delta_n$ values for the K-L index and INFO are presented in each of these figures. It should be noticed that the vertical scales for the plots presented in Figures 2 and 3 are not the same. The different vertical scales used here provide a better comparison for the different indices so that the differences among the indices can be clearly observed. It can be seen in Figure 2 that the K-L index having different $\delta_n$ values have similar or smaller bias in estimating the true $\theta$s than INFO. It appears that the K-L index having $\delta_n = 3/\sqrt{n}$ outperforms the K-L indices having other $\delta_n$s. In addition, the K-L index having $\delta_n = 3/\sqrt{n}$ has much smaller bias than INFO when $\theta = -3.0$ and $\theta = 2.0$. It also has a much smaller MSE than INFO at four out of six true $\theta$ values: $\theta = -3.0$, $\theta = -2.0$, $\theta = -1.0$, and $\theta = 2.0$. It has similar MSE as INFO at the
other two true $\theta$ values. Figure 1 illustrates that $\delta_n = 3/\sqrt{n}$ has the slowest convergence rate of the four $\delta_n$ values studied. In other words, the transformation from K-L to INFO for this index is much slower than for the other K-L indices. The better performance in terms of bias and MSE for this index may be due to this slow transformation from K-L to INFO.

When $\theta_0 = 3.0$, the bias and MSE statistics slightly increase when more items are selected for all the indices studied. One explanation might be that the total number of items selected is too small ($n = 20$) to produce a stable estimate at $\theta_0 = 3.0$. This phenomenon was not observed in the simulated data case when the number of items selected was 30.

**Simulated Data**

The comparisons between the K-L and INFO using simulated data are summarized in Figures 4 and 5. Because results from the real data pool analyses indicate that the K-L having $\delta_n = 3/\sqrt{n}$ produced the best $\theta$ estimates among the four K-L indices, only the K-L index having $\delta_n = 3/\sqrt{n}$ were used to select items from simulated item pool. All plots, except the last one, in Figure 4 share a common vertical scale to enable the comparison of the performance of these two indices at different true $\theta$ values. The vertical scale for $\theta_0 = 3.0$ is larger than the other five plots in Figure 4 to cover the extreme values. All plots in Figure 5 are on a common scale.

When $\theta_0 = -3.0$, INFO has smaller bias and MSE than K-L. When true $\theta$ is between -2.0 and 1.0, K-L outperforms INFO in terms of both bias and MSE. When $\theta \geq 2.0$, the performance of the two indices are very similar.

The performance of INFO and K-L when 10, 15, 20, 25, or 30 items are selected can also be compared across $-3.0 < \theta_0 < 3.0$. It can be seen in
Figures 6 and 7 that K-L has smaller bias and MSE when \(-2.0 < \theta_o < 2.0\) and when the length of CAT ranges from 10 to 30.

Conclusions

The results of this study, based on both simulated and real data, provides evidence that Chang and Ying's global information method produces similar or better \(\theta\) estimates than the more traditional information approach in CAT item selection. Because the present study is exploratory, further studies need to be carried out. For example, different \(\delta_n\) values in the K-L index need to be further compared using simulation technique to confirm the results found in this study, namely, that \(\delta_n = 3/Vn\) produces better results than other \(\delta_n\) values. In addition, the performance of K-L and INFO for other \(\theta_o\) values, such as \(\theta_o = -2.5, -1.5, 1.5,\) or 2.5 also need to be studied and compared.

In conclusion, the K-L index appears to be a very promising statistical index in CAT item selection. However, the K-L index is more computationally intense than INFO because integration is involved in the calculations. Another study that might be done would look at ways to make the K-L index more computationally efficient.

References


Table 1
Descriptive Statistics for the Real and the Simulated Item Pools
(N = 500)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>REAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1.29</td>
<td>0.44</td>
<td>0.20</td>
<td>2.35</td>
</tr>
<tr>
<td>B</td>
<td>0.17</td>
<td>0.66</td>
<td>-0.99</td>
<td>2.74</td>
</tr>
<tr>
<td>C</td>
<td>0.21</td>
<td>0.13</td>
<td>0.00</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>SIM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1.09</td>
<td>0.49</td>
<td>0.27</td>
<td>2.44</td>
</tr>
<tr>
<td>B</td>
<td>-0.02</td>
<td>1.55</td>
<td>-3.00</td>
<td>2.97</td>
</tr>
<tr>
<td>C</td>
<td>0.24</td>
<td>0.10</td>
<td>0.02</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Figure 1

The Four $\delta_n$ Sequences in the K-L Information Index
Figure 2 Bias at Six $\theta$, -- Real Data

BIAS, LENGTH=20
TRUE THETA=-3.0

- $3/\sqrt{n}$
- $1/n$
- $1/exp(0.1'n)$
- $3/exp(0.1'n)$
- info

BIAS, LENGTH=20
TRUE THETA=-2.0

- $3/\sqrt{n}$
- $1/n$
- $1/exp(0.1'n)$
- $3/exp(0.1'n)$
- info

BIAS, LENGTH=20
TRUE THETA=-1.0

- $3/\sqrt{n}$
- $1/n$
- $1/exp(0.1'n)$
- $3/exp(0.1'n)$
- info

BIAS, LENGTH=20
TRUE THETA=1.0

- $3/\sqrt{n}$
- $1/n$
- $1/exp(0.1'n)$
- $3/exp(0.1'n)$
- info

BIAS, LENGTH=20
TRUE THETA=2.0

- $3/\sqrt{n}$
- $1/n$
- $1/exp(0.1'n)$
- $3/exp(0.1'n)$
- info

BIAS, LENGTH=20
TRUE THETA=3.0

- $3/\sqrt{n}$
- $1/n$
- $1/exp(0.1'n)$
- $3/exp(0.1'n)$
- info
Figure 3  MSE at Six $\theta$ -- Real Data

MSE, LENGTH=20
TRUE THETA=-3.0

MSE, LENGTH=20
TRUE THETA=-2.0

MSE, LENGTH=20
TRUE THETA=-1.0

MSE, LENGTH=20
TRUE THETA=1.0

MSE, LENGTH=20
TRUE THETA=2.0

MSE, LENGTH=20
TRUE THETA=3.0
Figure 4  Bias at Six $\theta_0$ -- Simulated Data

Bias, LENGTH=30
TRUE THETA=-3.0

Bias, LENGTH=30
TRUE THETA=-2.0

Bias, LENGTH=30
TRUE THETA=-1.0

Bias, LENGTH=30
TRUE THETA= 1.0

Bias, LENGTH=30
TRUE THETA= 2.0

Bias, LENGTH=30
TRUE THETA= 3.0
Figure 5  MSE at Six $\theta_0$ -- Simulated Data

MSE, LENGTH=30
TRUE THETA= -3.0

MSE, LENGTH=30
TRUE THETA= -2.0

MSE, LENGTH=30
TRUE THETA= -1.0

MSE, LENGTH=30
TRUE THETA=  1.0

MSE, LENGTH=30
TRUE THETA=  2.0

MSE, LENGTH=30
TRUE THETA=  3.0
Figure 6  Bias at Different Number of Items Selected

BIAS, ITEM=10

BIAS, ITEM=15

BIAS, ITEM=20

BIAS, ITEM=25

BIAS, ITEM=30

Bias at Different Number of Items Selected

--- K-L --- INFO

--- K-L --- INFO

--- K-L --- INFO

--- K-L --- INFO

--- K-L --- INFO

--- K-L --- INFO

--- K-L --- INFO
Figure 7 MSE at Different Number of Items Selected

MSE, ITEM=10

MSE, ITEM=15

MSE, ITEM=20

MSE, ITEM=25

MSE, ITEM=30

BEST COPY AVAILABLE