DOCUMENT RESUME

ED 398 252 TM 025 201

AUTHOR	Tang, K. Linda
TITLE	A Comparison of the Traditional Maximum Information Method and the Global Information Method in CAT Item Selection.
PUB DATE	Apr 96
NOTE	21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Ability; *Adaptive Testing; Comparative Analysis; *Computer Assisted Testing; Estimation (Mathematics); *Selection; Test Construction; *Test Items
IDENTIFIERS	*Information Function (Tests)

ABSTRACT

The average Kullback-Keibler (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 δ_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 θ 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 θ 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 ($\hat{\theta}_n$) 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 $\hat{\theta}_n$ is not close to the examinee's true ability (θ_0), the item which has maximum local information at $\hat{\theta}_n$ may not be the most appropriate item to administer to the examinee having true ability θ_0. 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 (θ_0) 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 θ_0 from any other θ 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:

$$K_j(\hat{\theta}_n) = \int_{\theta_n-\delta_n}^{\theta_n+\delta_n} K_j(\theta \| \hat{\theta}_n)\, d\theta \ , \tag{1}$$

where $\hat{\theta}_n$ is the maximum likelihood estimator of θ_0 based on n items administered to an examinee, and $K_j\ (\theta || \hat{\theta}_n)$ is defined as

$$K_j(\theta \| \hat{\theta}_n) = P_j(\hat{\theta}_n) \log(\frac{P_j(\hat{\theta}_n)}{P_j(\theta)}) + (1-P_j(\hat{\theta}_n)) \log(\frac{1-P_j(\hat{\theta}_n)}{1-(P_j(\theta))}) \ . \tag{2}$$

The parameter δ_n in (1) generates a sequence which converges to 0 as n increases. The δ_n controls the width of the interval under the K-L information index curve. This interval is expected to contain the true ability parameter, θ_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 $\hat{\theta}_n$ approaches θ_0. As Chang and Ying pointed out, for a small δ_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 δ_n value, the area is very much influenced by the global information. Therefore, δ_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 δ_n values used in (1) were $2/\sqrt{n}$ and $1/n$. However, the effect of these two δ_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 δ_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 (LOG(a) ~ $N(0,0.5^2)$). 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 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.

Insert Table 1 about here

2. True abilities

The following six true abilities (θ_0s) were used in the study: -3.0, -2.0, -1.0, -1.0, 2.0, 3.0. These θ_0 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. δ_n used in the average K-L information index

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

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

Insert Figure 1 about here

Figure 1 illustrates that the four δ_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 δ_n value is small. The shifting of K-L to the traditional information occurs when the δ_n sequence approaches to 0. In other words, the K-L indices associated with each of the four δ_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 $\delta_n = 1/e^{0.1*n}$ shifts from global to local information around the administration of the 25th item; and the K-L index having $\delta_n = 3/e^{0.1*n}$ shifts from global to local information ar‹ ᵔnd the administration of the 35 ıtem. The K-L index having $\delta_n = 3/\sqrt{n}$ clearly converges to 0 much slowly than the other three δ_n values.

In summary, there were 30 experimental conditions in the real data case: 6 true θ values (-3.0, -2.0, -1.0, 1.0, 2.0, and 3.0) x 4 δ_n values ($3/\sqrt{n}$, $1/n$, and $1/e^{0.1*n}$, and $3/e^{0.1*n}$) for the global information method plus the 6 true θ 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 θ 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 θ 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 θ estimate after item i is administered to examinee j having true ability θ_0 is defined as

$$B_i = \frac{1}{100}\sum_{j=1}^{100}\theta_{ij} - \theta_0 \ , \qquad (3)$$

and mean squared error for the θ estimate after item i is administered to examinee j having true ability θ_0 is defined as

$$MSE_i = \frac{1}{100}\sum_{j=1}^{100} \ (\theta_{ij} - \theta_0)^2 \ , \qquad (4)$$

where $\hat{\theta}_{ij}$ is the estimate for θ_0 obtained using either the global information method or the conventional information method, and i = 3,..., 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 θ values for the K-L index and the conventional information index (INFO). Each of the four δ_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 δ_n values have similar or smaller bias in estimating the true θs than INFO. It appears that the K-L index having $\delta_n = 3/\sqrt{n}$ outperforms the K-L indices having other δ_ns. 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 θ 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 θ values. Figure 1 illustrates that $\delta_n = 3/\sqrt{n}$ has the slowest convergence rate of the four δ_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 θ 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 θ 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 θ 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 $-2.0 < \theta_0 < 2.0$ 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 θ 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 δ_n values in the K-L index need to be further compared using simulation technique to confirm the results fond in this study, namely, that $\delta_n = 3/\sqrt{n}$ produces better results than other δ_n values. In addition, the performance of K-L and INFO for other θ_0 values, such as $\theta_0 = -2.5, -1.5, 1.5$, 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

Chang, H. (1995, April). A global information approach to computer adaptive testing. Paper presented at the annual meeting of the NCME, San Francisco.

Chang, H., & Ying, Z. (in press). A global information approach to computerized adaptive testing. Applied Psychological Measurement.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Chicago,. IL: Scientific Software, INC.

Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 31, 3-31.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mood, A. M., Graybill, F. A., & Boes, D. (1985). Introduction to the theory of statistics (3rd Edition). New York: McGraw-Hill Book Company.

Table 1

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

		Mean	SD	Min	Max
REAL	A	1.29	0.44	0.20	2.35
	B	0.17	0.66	-0.99	2.74
	C	0.21	0.13	0.00	0.70
SIM	A	1.09	0.49	0.27	2.44
	B	-0.02	1.55	-3.00	2.97
	C	0.24	0.10	0.02	0.62

Figure 1

The Four δ_n Sequences in the K-L Information Index

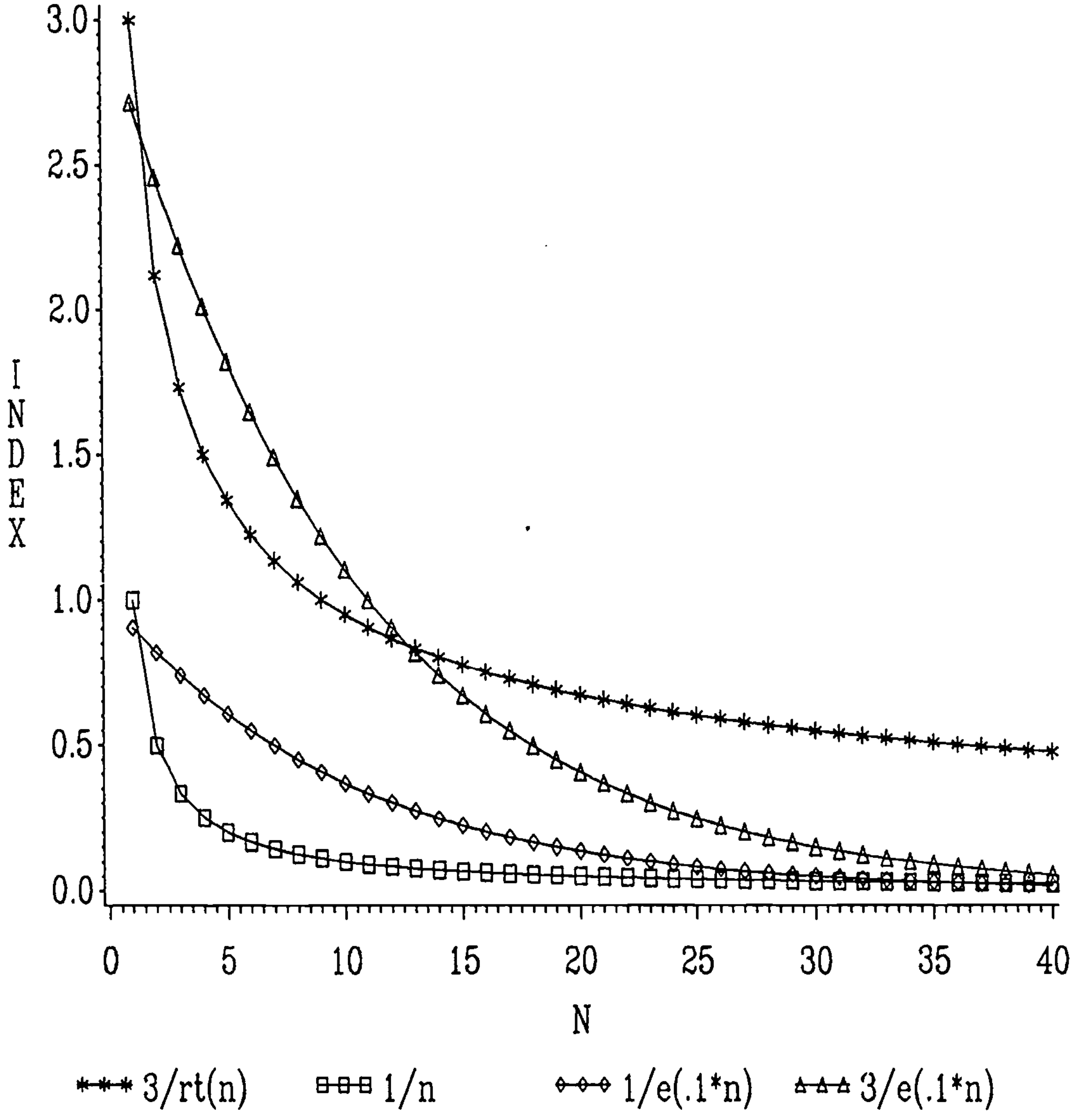


Figure 2 Bias at Six θ_0 -- Real Data

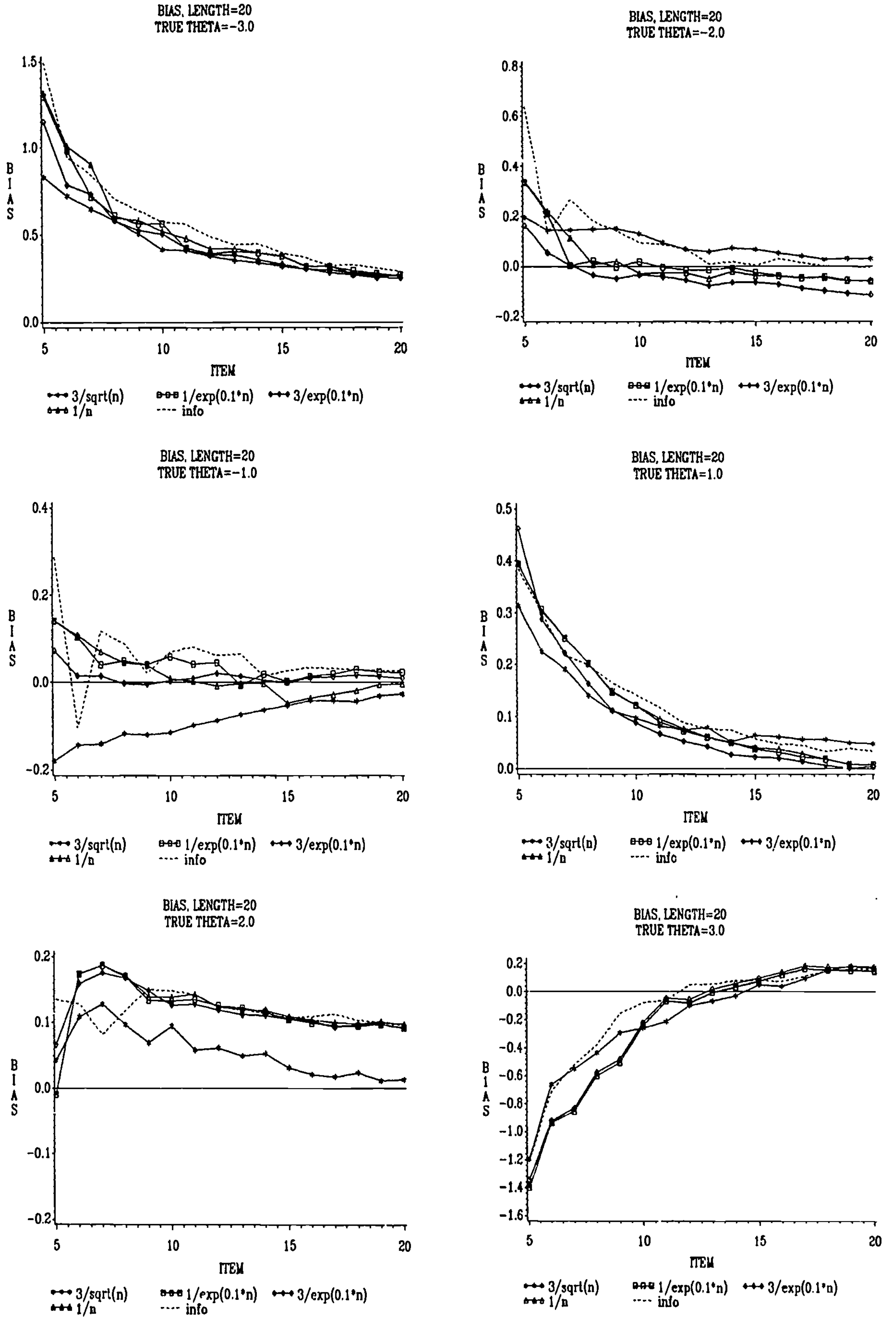


Figure 3 MSE at Six θ_0 -- Real Data

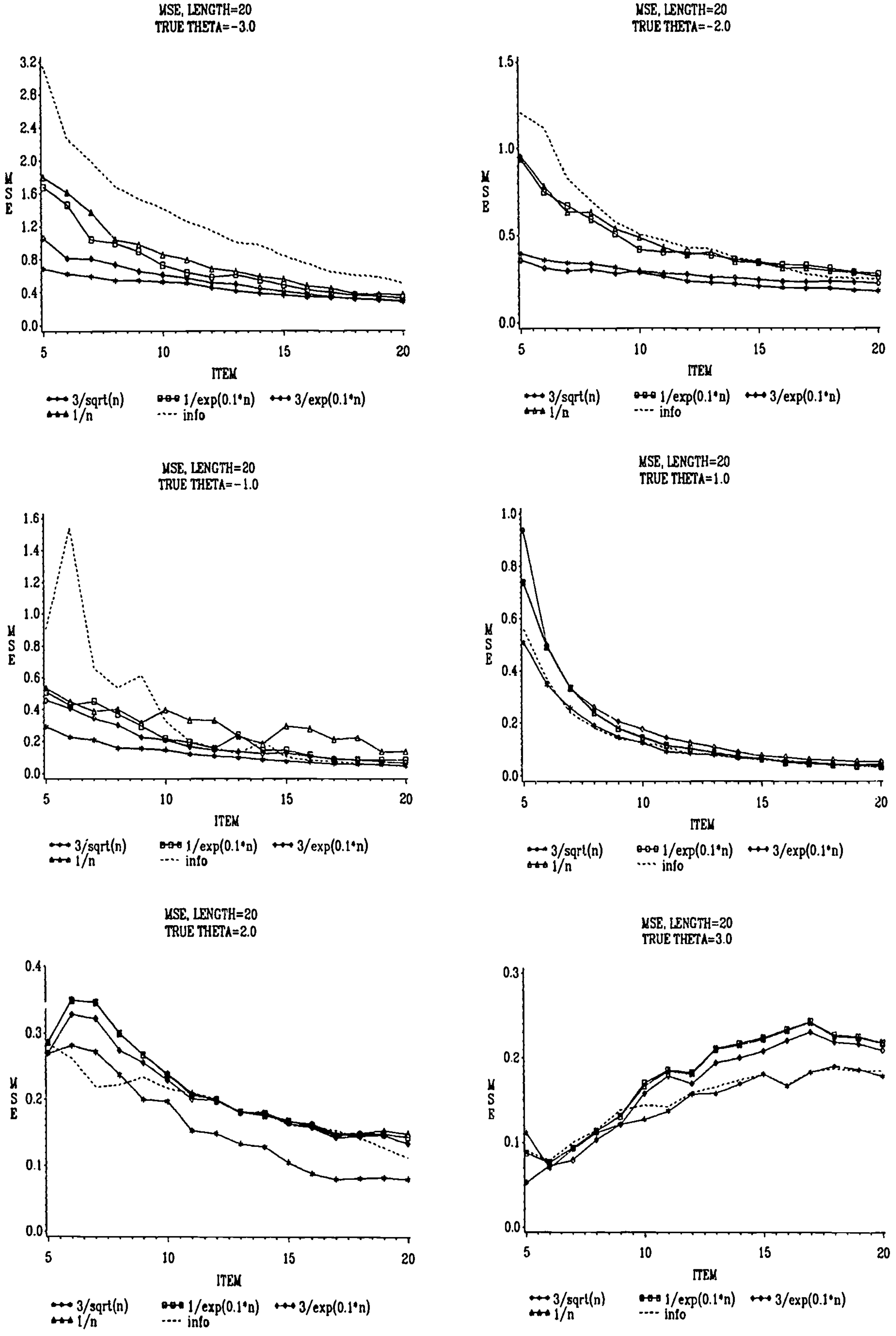


Figure 4 Bias at Six θ_0 -- Simulated Data

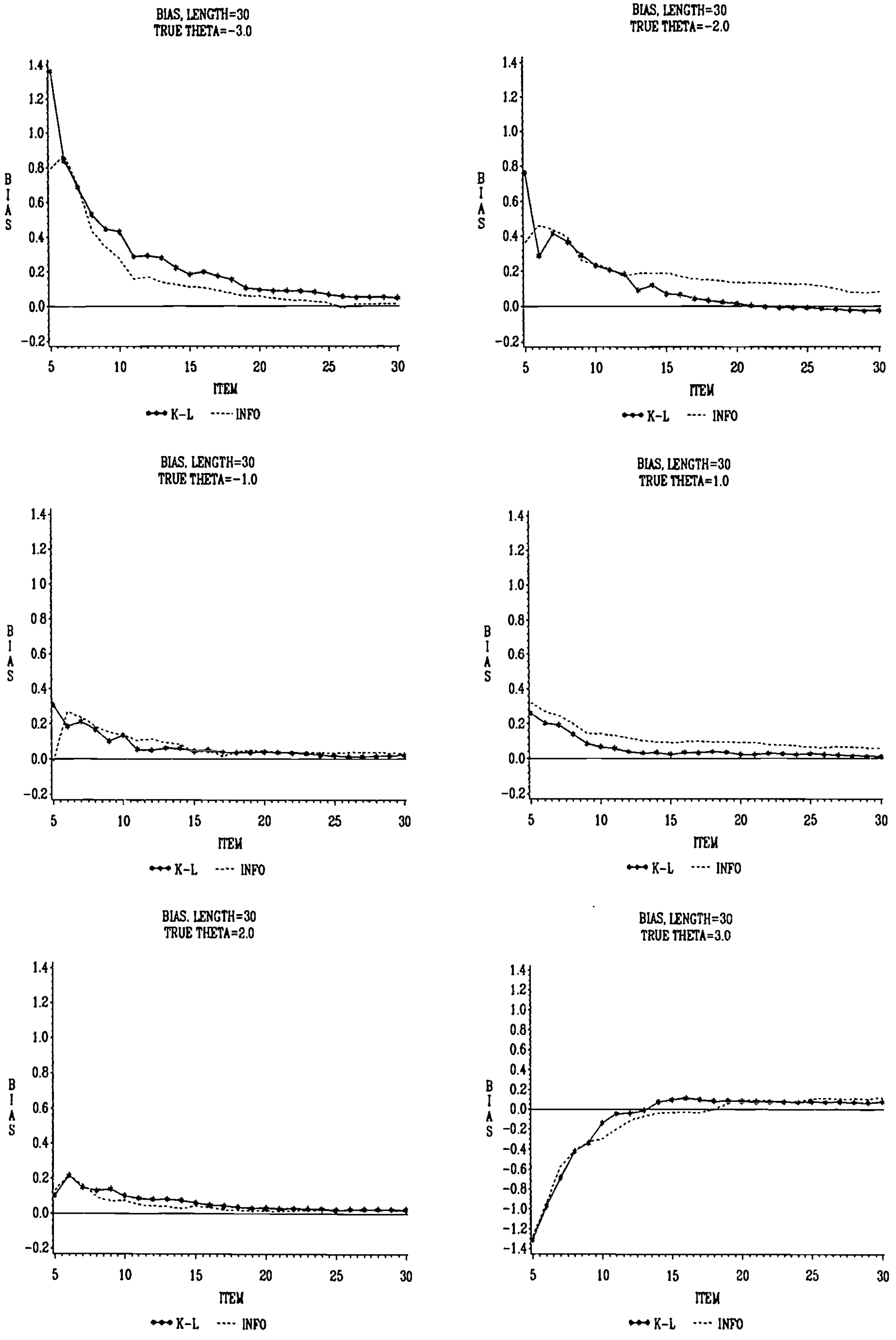


Figure 5 MSE at Six θ_0 -- Simulated Data

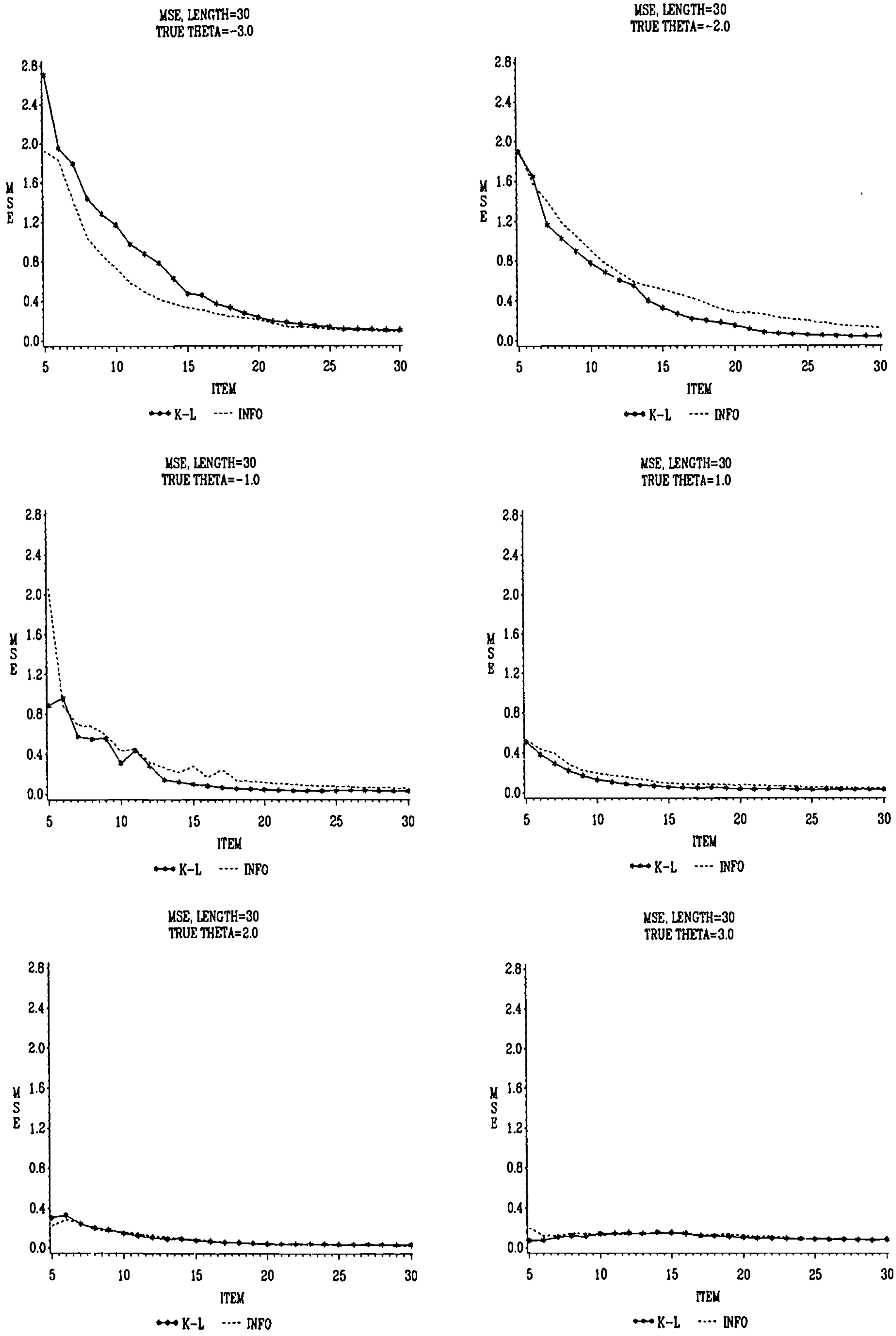


Figure 6 Bias at Different Number of Items Selected

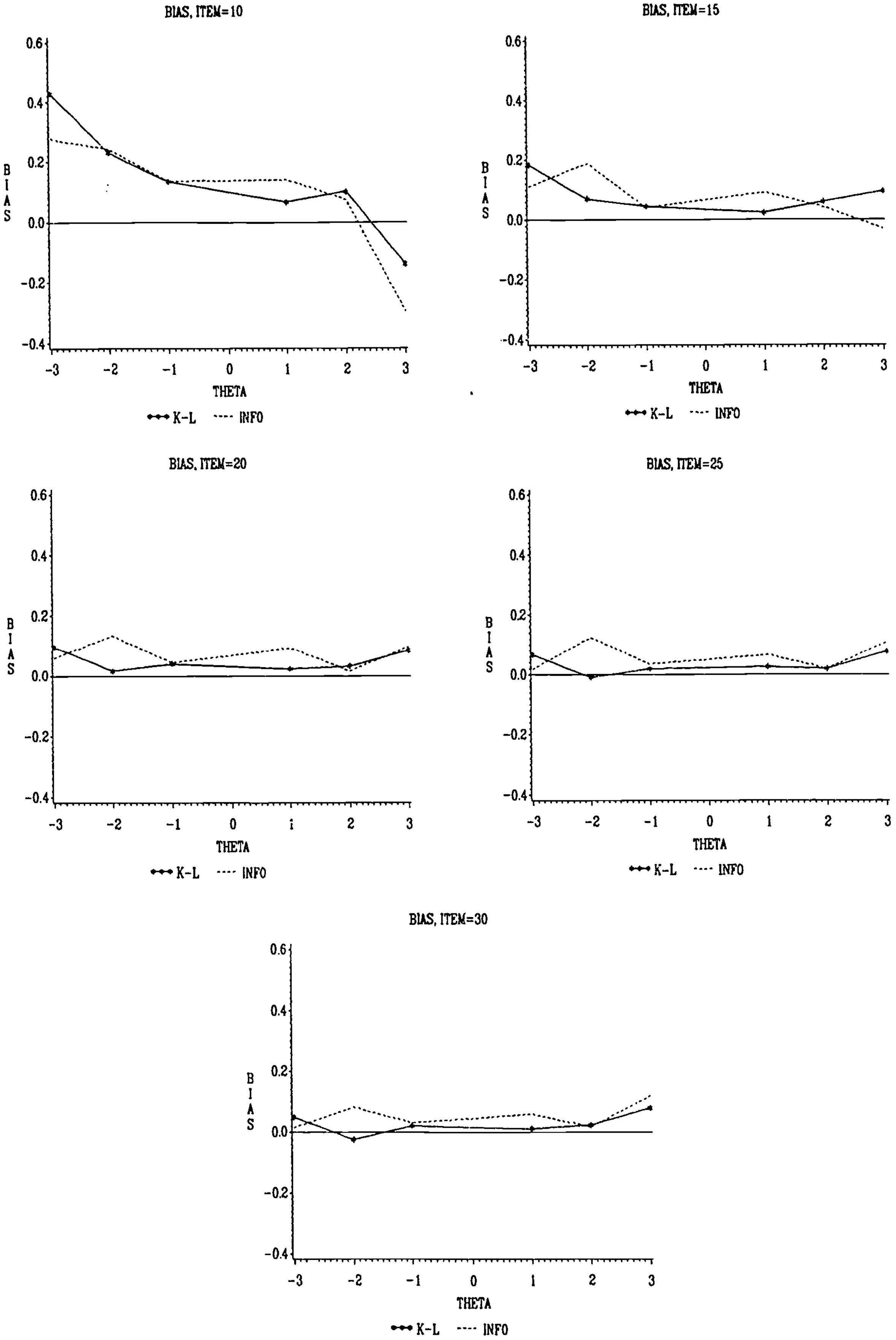
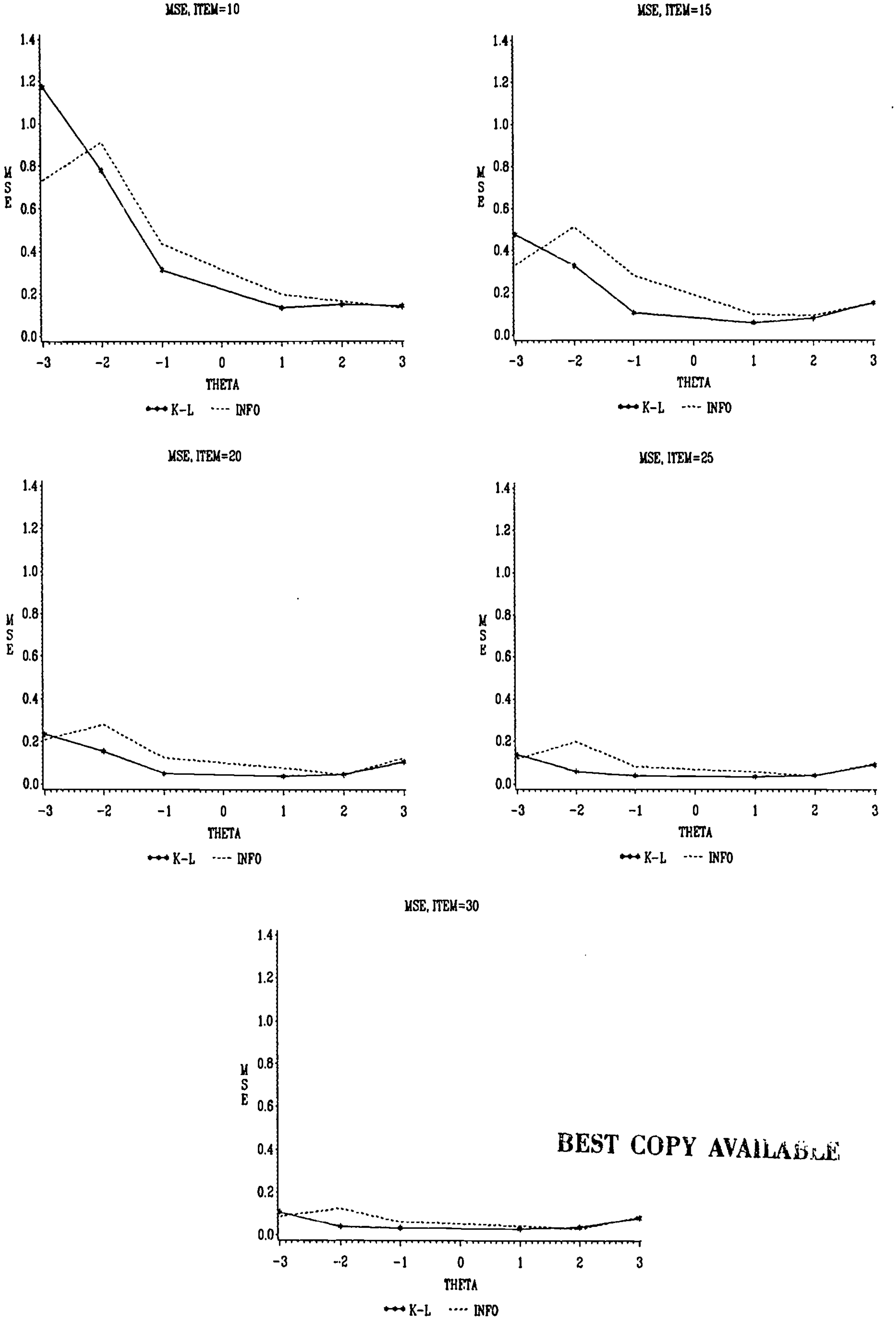


Figure 7 MSE at Different Number of Items Selected



MSE, ITEM=10

MSE, ITEM=15

MSE, ITEM=20

MSE, ITEM=25

MSE, ITEM=30