A Global Information Approach to Computerized Adaptive Testing

Hua-Hua Chang, Educational Testing Service

Zhiliang Ying, Rutgers University

Most item selection in computerized adaptive testing is based on Fisher information (or item information). At each stage, an item is selected to maximize the Fisher information at the currently estimated trait level (θ). However, this application of Fisher information could be much less efficient than assumed if the estimators are not close to the true θ, especially at early stages of an adaptive test when the test length (number of items) is too short to provide an accurate estimate for true θ. It is argued here that selection procedures based on global information should be used, at least at early stages of a test when θ estimates are not likely to be close to the true θ. For this purpose, an item selection procedure based on average global information is proposed. Results from pilot simulation studies comparing the usual maximum item information item selection with the proposed global information approach are reported, indicating that the new method leads to improvement in terms of bias and mean squared error reduction under many circumstances. *Index terms: computerized adaptive testing, Fisher information, global information, information surface, item information, item response theory, Kullback-Leibler information, local information, test information.*

Computerized adaptive testing (CAT) was proposed by Lord (1971), Owen (1975), and Weiss (1976), among others, to measure the trait levels (θs) of examinees with greater precision than conventional tests by building an individualized test for each examinee. Test items are selected sequentially, according to the current performance of an examinee. The test is tailored to each examinee's θ level, thus matching the difficulties of the items to the examinee being measured. Able examinees can avoid responding to too many easy items, and less able examinees can avoid being exposed to too many difficult items. The major advantage of CAT is that it provides more efficient trait estimates with fewer items than that required in conventional tests (e.g., Weiss, 1982). Significant progress has been made in the development and implementation of CAT due, in part, to the rapid advancement of computer technology (Wainer, 1990). However, methodological as well as theoretical developments in CAT appear to be rather limited.

A basic ingredient in CAT is the item selection procedure that is used to select items during the course of the test. For the past two decades, the most commonly used item selection procedure has been based on maximizing item information. More specifically, an item is selected that has maximum information at the currently estimated θ level ($\hat{\theta}$), which is estimated from the available responses at that time. An alternative to the maximum information approach is the Bayesian method (e.g., Owen, 1975). Instead of using item information at $\hat{\theta}$, the Bayesian approach uses the posterior variance as the criterion for item selection. At the initial stages, posterior distributions depend heavily on the choice of prior distribution for θ, but the dependency diminishes at the later stages. Furthermore, according to Chang & Stout (1993), the posterior variance approaches the reciprocal of the test information when the number of items becomes large.

Item information typically has been defined as Fisher information, which varies from examinee to examinee and therefore is a function of θ. The value of Fisher information at the true θ level of a particular examinee, denoted by θ_0, indicates the efficiency of the item for estimation of θ. However, its value at a θ level

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 3, September 1996, pp. 213–229

0146-6216/96/030213-17$2.10

distant from θ_0 may not be a good indicator of efficiency. Because it uses Fisher information at the current $\hat{\theta}$ level, the information criterion can be inefficient if $\hat{\theta}$ is not close to θ_0. This may well be the case at early stages of a CAT when there are only a small number of items (providing little information) to construct a reliable estimator for θ. Consequently, items selected at an early stage may not be an efficient choice. For this reason, the issue of selecting "best" items at early stages has attracted much attention recently (e.g., Davey & Parshall, 1995; Fan & Hsu, 1995; Stocking, 1993; van der Linden, 1995; Veerkamp & Berger, 1994).

New methods and suggestions, along with theoretical and empirical studies, have been proposed to overcome inefficiency due to inaccurate estimation of θ_0. In particular, Veerkamp & Berger (1994) proposed an "interval information criterion": Instead of the item information at a point, their selection procedure is based on the highest mean value of the information function in a confidence interval [see Chang & Ying (1996b) for a discussion of Veerkamp and Berger's proposal]. However, Stocking (1993) argued that, in addition to item information, item selection should incorporate some further criteria, such as conditional and absolute exposure rates, item pool refreshment or replacement, test specifications, and item-type ordering.

It appears that further progress, if there is to be any, in the foundational research of CAT could occur in the area of item selection procedures. The usual large-sample results, such as consistency and asymptotic posterior normality, have been established for item response theory (IRT) models. Under general regularity conditions, these results ensure that commonly used estimators converge to θ_0. It then follows that the item information criterion described above should be close to optimal at later stages in a CAT when the number of administered items is already sufficiently large. Note that a major goal of CAT is to more efficiently estimate θ with fewer items. Reducing the number of items used in the test thus makes the quality of item selection at early stages extremely crucial. Therefore, developing necessary concepts and methods for small-sample selection becomes very important.

This paper presents (1) a new concept of information—global information and related information functions—that provides information when the estimator is not close to the true parameter; (2) an item selection procedure based on average global information; and (3) some results from a pilot simulation study comparing the standard maximum information approach with the new global information approach.

Fisher Information in IRT Models

The item response function for the ith item will be denoted by P_i and its complement by $Q_i = 1 - P_i$. Thus, an examinee with trait parameter θ will answer the item correctly with probability $P_i(\theta)$ and incorrectly with probability $Q_i(\theta)$. Following Lord (1980), the Fisher item information function is defined as

$$I_i(\theta) = \left[\frac{\partial P_i(\theta)}{\partial \theta}\right]^2 \Big/ P_i(\theta)Q_i(\theta). \tag{1}$$

For a test consisting of items $i = 1, \ldots, n$, the test information, as a function of θ, is simply the sum of the individual item information functions:

$$I^{(n)}(\theta) = \sum_{i=1}^{n} I_i(\theta). \tag{2}$$

Fisher information is closely related to maximum likelihood (ML) estimation. If an examinee's item responses are denoted generically by $X_1, \ldots, X_n$, the likelihood function can be written as

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} \left[P_i^{X_i}(\theta)Q_i^{1-X_i}(\theta)\right]. \tag{3}$$

Denote the log-likelihood function as

$$l_n(\theta) = \log L(\theta; X_1, \ldots, X_n). \tag{4}$$

$\hat{\theta}_n$ is used to denote the ML estimator

$$L(\hat{\theta}_n; X_1, \ldots, X_n) = \max_{\theta} L(\theta; X_1, \ldots, X_n), \tag{5}$$

or equivalently,

$$l_n(\hat{\theta}_n) = \max_{\theta} l_n(\theta). \tag{6}$$

Recall that θ_0 is true θ. The response to the ith item, X_i, is a random variable with probability mass function

$$P_i(\theta)^{x_i}[1 - P_i(\theta)]^{1-x_i} \quad x_i = 0, 1, \tag{7}$$

where x_i denotes the value taken by X_i.

The asymptotic variance of $\hat{\theta}_n$ is the reciprocal of $I^{(n)}(\theta_0)$, the test information function (Lehmann, 1983, p. 465). In other words, the Fisher information is inversely proportional to the error of the ML estimator.

When a CAT is administered to an examinee, a series of item selection decisions are made, each of which depends on the examinee's responses to the preceding items using the current $\hat{\theta}$, usually $\hat{\theta}_m$ if m items have been answered. An information-based sequential decision rule is to select the next item so that the information at $\hat{\theta}_m$ is maximized. Apparently, the appropriateness of this rule may depend on how close $\hat{\theta}_m$ is to θ_0. The item selection procedure typically used is based on I, which is reasonable when $\hat{\theta}_m$ is close to θ_0. However, if $\hat{\theta}_m$ is not close to θ_0, then I at $\hat{\theta}_m$ may not reflect the true information of the item. The deviation of $\hat{\theta}_m$ from θ_0 is likely to be non-negligible when m is small (i.e., at early stages of a CAT). In addition, I is relatively unstable for commonly used IRT models, including the three-parameter logistic model (3PLM; Chang & Ying, 1996a). It is also questionable whether a univariate function of θ alone is sufficient to capture the entire information content of an item. A more flexible approach may be needed for this problem.

Local Information

If the information around a small region of θ_0 is viewed as local information, then the information outside that region can be viewed as global information. In statistical testing theory, there are two kinds of alternatives to the null hypothesis—local and fixed. For example, if the null hypothesis is H_0: $\theta = \theta_0$, then a fixed alternative could be H_1: $\theta = \theta_1$, and local alternatives, relative to a sample of size m, could be H_1: $\theta = \theta_0 + (\theta_1/\sqrt{m})$. The local alternatives approach the null hypothesis as m increases, whereas the fixed alternative does not. It is reasonable to expect that local information would be related to the power of detecting local alternatives, and global information to that for a fixed alternative. With respect to CAT, local information may serve as a benchmark for item selection when there is sufficient knowledge about the location of θ_0, and global information might be preferred when there is lack of such knowledge.

In practice, θ_0 is unknown. For an information-based criterion to be useful, the value of information at every possible θ has to be specified. When information is defined for every θ, it effectively becomes a function on the entire parameter space. The local information (function) should then mean that at each θ, its value measures the amount of information the item contains when the examinee's true but unknown trait level is θ.

Test Information is Local Information

$I^{(n)}$ in IRT represents a local information function. Recall that $I^{(n)}$ for a given item response sequence X_1, ..., X_n can be written as

$$I^{(n)}(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial}{\partial\theta}\log P_i(X_i|\theta)\right]^2 = \mathrm{E}\left[\frac{\partial}{\partial\theta}\log L(\theta; X_1, \ldots, X_n)\right]^2. \quad (8)$$

Because $I^{(n)}$ in IRT is defined as the Fisher information, its meaning and justification may be described by paraphrasing Lehmann (1983, p. 117):

> The function
>
> $$\frac{\frac{\partial}{\partial\theta} L(\theta; x_1, \ldots, x_n)}{L(\theta; x_1, \ldots, x_n)} \quad [(9)]$$
>
> is the relative rate at which the density changes at x_1, ..., x_n. The average of the square of this rate is expressed by Equation 8. It is plausible that the greater this expectation is at a given value θ_0, the easier it is to distinguish θ_0 from neighboring values θ, and, therefore, the more accurately θ can be estimated at $\theta = \theta_0$. The quantity $I^{(n)}(\theta)$ is called the information or the Fisher information that X_1, ..., X_n contains about parameter θ.

Lehmann (1983) emphasized that "...the surmise turns out to be correct when sample size is large" (p. 118). The asymptotic theory of Le Cam (Le Cam & Yang, 1990) provides quantification in terms of statistical hypothesis testing for local alternatives.

Suppose a test with n items is to be designed to estimate θ_0. According to Hambleton & Swaminathan (1985), I "...can be interpreted as providing per unit discrimination between ability levels" (p. 102) that are close together. This implies that for any fixed individual with θ_0, I is the discrimination power between θ_0 and any θ_1 that is close to θ_0. Thus, for any fixed θ_0, I is the local information that the item contains about θ_0.

Let $\hat{\theta}_n$ denote the ML estimator or its asymptotically equivalent variant. It is important that items be selected to make $\hat{\theta}_n$ as close as possible to θ_0. As n increases, $\hat{\theta}_n$ approaches θ_0; in fact, it is asymptotically normal with mean θ_0 and variance $1/I^{(n)}(\theta_0)$. The closeness of $\hat{\theta}_n$ to θ_0 is thus governed by $I^{(n)}(\theta_0)$: the larger $I^{(n)}(\theta_0)$ is, the closer $\hat{\theta}_n$ is to θ_0. Thus, provided n is large, an efficient test may be obtained by making $I^{(n)}(\theta_0)$ as large as possible.

However, for small ns, the estimator may not be close to θ_0, in which case the information inside a small region around $\hat{\theta}_n$ would not be useful. The term "information function" may be misleading if it is used without considering its asymptotic properties (Lord, 1971, p. 10). Thus, global information for the situation in which $\hat{\theta}_n$ is not close to θ_0 is needed.

Global Information

Given an examinee's responses X_1, ..., X_n to the n items in a test, the quantity that summarizes all the information for the examinee's θ is the likelihood function $L(\theta) = L(\theta; X_1, \ldots, X_n)$ defined by Equation 3. To distinguish any fixed θ_1 from θ_0, examine the difference between values of L at θ_1 and θ_0. Such a difference can be captured by the ratio of the two values, resulting in the well-known likelihood ratio test (Neyman & Pearson, 1936). By Neyman-Pearson theory (Lehmann, 1986), the likelihood ratio method is optimal for testing $\theta = \theta_0$ versus $\theta = \theta_1$. In other words, it is the best way to tell θ_1 from θ_0 when the IRT model is assumed for X_1, ..., X_n observed.

Because the errors associated with the likelihood ratio test decrease to 0 exponentially fast (Serfling, 1980, §10.3.2), it is convenient to take the logarithm of the likelihood ratio. Moreover, according to Lehmann (personal communication, September 1, 1995), one of the main reasons for taking the logarithm is that the likelihood is a product, but its logarithm is a sum, which is much easier to work with. One of the consequences is the additivity of information that would not be possible without taking logs. The expected value of the log-likelihood ratio quantifies how powerful (efficient) the statistical test is and is commonly known as the Kullback-

Leibler (KL) information (Cover & Thomas, 1991; Kullback, 1959). It also measures the discrepancy between the two probability distributions specified by θ_0 and θ_1.

Kullback-Leibler Information

Definition 2.1: KL item information. Let θ_0 be the true parameter. For any θ, the KL information of the ith item (with response X_i) is defined by

$$K_i(\theta\|\theta_0) \equiv \mathrm{E}_{\theta_0} \log\left[\frac{L_i(\theta_0; X_i)}{L_i(\theta; X_i)}\right], \tag{10}$$

where E_{θ_0} denotes expectation over X_i and

$$L_i(\theta; X_i) = P_i^{X_i}(\theta) Q_i^{1-X_i}(\theta) \tag{11}$$

is the likelihood function for the ith item.

A straightforward probability calculation using Equation 7 shows that the item KL information can be expressed explicitly as

$$K_i(\theta\|\theta_0) = P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + \left[1 - P_i(\theta_0)\right]\log\left[\frac{1-P_i(\theta_0)}{1-P_i(\theta)}\right]. \tag{12}$$

[The notation of double vertical bars is standard for KL information (Cover & Thomas, 1991, p. 18). The double bars, which signify that θ needs to be separated from θ_0, are used to avoid confusion with the single bar, which typically indicates conditioning.]

Note that as a function of θ and θ_0, K_i is not symmetric [i.e., $K_i(\theta\|\theta_0) \neq K_i(\theta_0\|\theta)$]. Furthermore, $K_i(\theta\|\theta_0) \geq 0$ and $K_i(\theta_0\|\theta_0) = 0$. Mimicking $I^{(n)}$, the corresponding KL test information can be defined.

Definition 2.2: KL test information. Let θ_0 be the true parameter. For any θ, the KL information for a test is defined by

$$K^{(n)}(\theta\|\theta_0) \equiv \mathrm{E}_{\theta_0} \log\left[\frac{L(\theta_0; X_1, \ldots, X_n)}{L(\theta; X_1, \ldots, X_n)}\right], \tag{13}$$

where $X_1, \ldots, X_n$ are the scored responses.

Note again that the expectation is with respect to $(X_1, \ldots, X_n)$. From this definition it follows that

$$K^{(n)}(\theta\|\theta_0) = \mathrm{E}_{\theta_0}\left[l_n(\theta_0) - l_n(\theta)\right] = \sum_{i=1}^{n} K_i(\theta\|\theta_0). \tag{14}$$

Again $K^{(n)}(\theta\|\theta_0) \geq 0$, and it is equal if $\theta = \theta_0$. K is sometimes referred to as the relative entropy or the KL distance (Cover & Thomas, 1991).

Analogous to $I^{(n)}$ defined by Equation 8, an important feature of $K^{(n)}$ is that the contribution of each item to the total information is additive. Thus, the total amount of information for a test can be readily determined. This feature is highly desirable in CATs because it enables test developers to separately calculate the information for each item and combine them to form updated test information at each stage.

Another important feature is that K is a function of two levels, θ and θ_0. K represents the discrimination power of the item on the two levels. It does not require that θ be close to θ_0. In this sense, K summarizes information content of the item with respect to a broad spectrum of θ levels. In contrast, I is a function of θ_0 only and represents discrimination power around θ_0 (Hambleton & Swaminathan, 1985, p. 102).

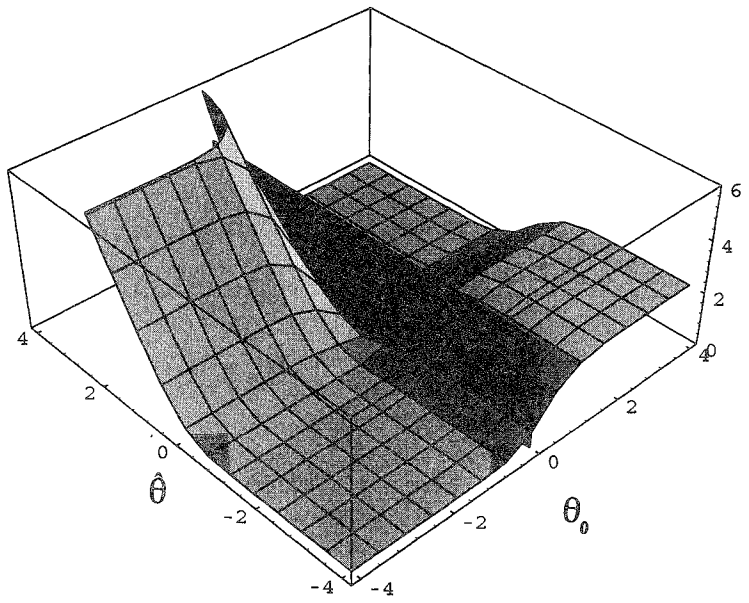
KL Information Is Global Information

The purpose of a CAT is to accurately estimate an examinee's θ_0 by efficiently selecting items. To this end,

it is desirable to find a quantity that distinguishes all those $\theta \neq \theta_0$ from θ_0. As argued above, the log-likelihood ratio is, in a sense, the best quantity constructed from data that can be used to distinguish θ from θ_0. K is the average (expectation) of the log-likelihood ratio. For item i, as θ varies over the parameter space, K generates a global profile about the discrimination power of the item. There is no requirement that θ be close to θ_0. In this sense, K may be viewed as a way to quantify global information.

For each θ_0, K is a function of θ, and I is a fixed number. This is one of the key distinctions between K and I. If θ_0 is allowed to vary across the entire scale, K becomes a global information surface in a three-dimensional space $(\lambda,\upsilon,\kappa)$, with λ corresponding to θ_0, υ to $\hat{\theta}$, and κ to K_i (see Figure 1). Figure 1 shows the KL information surface intersected with a vertical plane at $\lambda = 0$, for an item with 3PLM parameters [a (discrimination), b (difficulty), and c (pseudoguessing)]. The resulting curve on the plane is the KL information function at $\theta_0 = 0$. The geometrical meaning of a KL information function for a fixed θ_0 is a curve, which represents the intersection of the vertical plane $\lambda = \theta_0$ and the information surface. From Figure 1, observe that the KL information function changes its shape as θ_0 changes its values. No matter how it changes, K is always 0 along the entire 45° line ($\hat{\theta} = \theta_0$). Note that the curvature at $\theta_0 = \hat{\theta}$ equals I at θ_0.

Figure 1
KL Information Surface for an Item With $a = 3.0$, $b = 0.0$, $c = .1$, Intersected With a Vertical Plane $\lambda = 0$



Use of $K^{(n)}$ (or K) as a global characteristic is not new. In addition to the above discussion of its use in testing statistical hypotheses (the likelihood ratio test), statistical estimation is another use. In theoretical development of ML estimation, two basic properties—consistency and asymptotic normality—are commonly investigated. To establish consistency, the behavior of the likelihood function in the entire parameter space is examined to show that the values not close to the true parameter are not likely to maximize the likelihood function. This is often accomplished using $K^{(n)}$. To be more specific, it is expected that for θ distant from θ_0,

$L(\theta; X) < L(\theta_0; X)$ or equivalently $l_n(\theta; X) < l_n(\theta_0; X)$. $El_n(\theta; X) < El_n(\theta_0; X)$ is used to rule out those θ that are not close to θ_0. Having shown the consistency, the likelihood function in a region close to θ_0 is examined and asymptotic normality is established, which is connected to $I^{(n)}$. In this context, $K^{(n)}$ is used to study the likelihood function on the entire parameter space, whereas $I^{(n)}$ is only used locally around the true parameter.

The Relationship Between Local and Global Information

Global information should be used when n is small, and local information should be used when n is large. Thus, to design a good CAT, both global and local information are needed at different stages of the test. A practical problem from this consideration is how to determine the cut-off point and whether a smoother transition is needed. Moreover, it would certainly be desirable if a single measure could be constructed that mimics global information with small n, and local information with larger n. Thus, a connection between the local and the global information functions must be established.

Recall that for a person with $\theta = \theta_0$, $K^{(n)}(\theta\|\theta_0)$ is minimized at $\theta = \theta_0$ with minimum value $K^{(n)}(\theta\|\theta_0) = 0$. Thus, the derivative of $K^{(n)}(\theta\|\theta_0)$ at $\theta = \theta_0$ must be 0:

$$\frac{\partial}{\partial\theta} K^{(n)}(\theta\|\theta_0)\Big|_{\theta=\theta_0} = 0. \tag{15}$$

Through its Taylor series expansion at θ_0, the local variation of $K^{(n)}(\theta\|\theta_0)$ is then characterized primarily by its second derivative, which, not surprisingly, is $I^{(n)}$. More precisely,

$$\frac{\partial^2}{\partial\theta^2} K^{(n)}(\theta\|\theta_0)\Big|_{\theta=\theta_0} = I^{(n)}(\theta_0). \tag{16}$$

For any θ, $K^{(n)}$ represents the ease or difficulty of distinguishing θ from θ_0. In particular, for θ varying around θ_0, it also gives local information, which is connected to $I^{(n)}$. Equation 16 is simply a mathematical statement about this. Geometrically speaking, if K is viewed as a curve on the plane, I becomes the curvature of the curve at $\theta = \theta_0$. Note that both Equations 15 and 16 hold with $K^{(n)}$ replaced by K and $I^{(n)}$ by I.

Figure 2 plots KL information functions for five items with $\theta_0 = 1$. For each function, the curvature at 1 is equal to the value of I at 1. All the well-known influences, such as guessing and discrimination (Lord & Novick, 1968, pp. 460–464), on I have corresponding effects on the curvature of the KL functions. Note that in terms of I, Item 5 provides more information than Item 4; however, this is not the case for K, which shows that their relationship is more complex.

Figure 3 plots both K and I for two items at $\theta_0 = 0$. Although I for Item 1 is greater than that for Item 2 around $\theta_0 = 0$, it appears that Item 2 might be a better choice based on K, which shows that Item 2 is more "robust" and has more overall power when considering the entire parameter range.

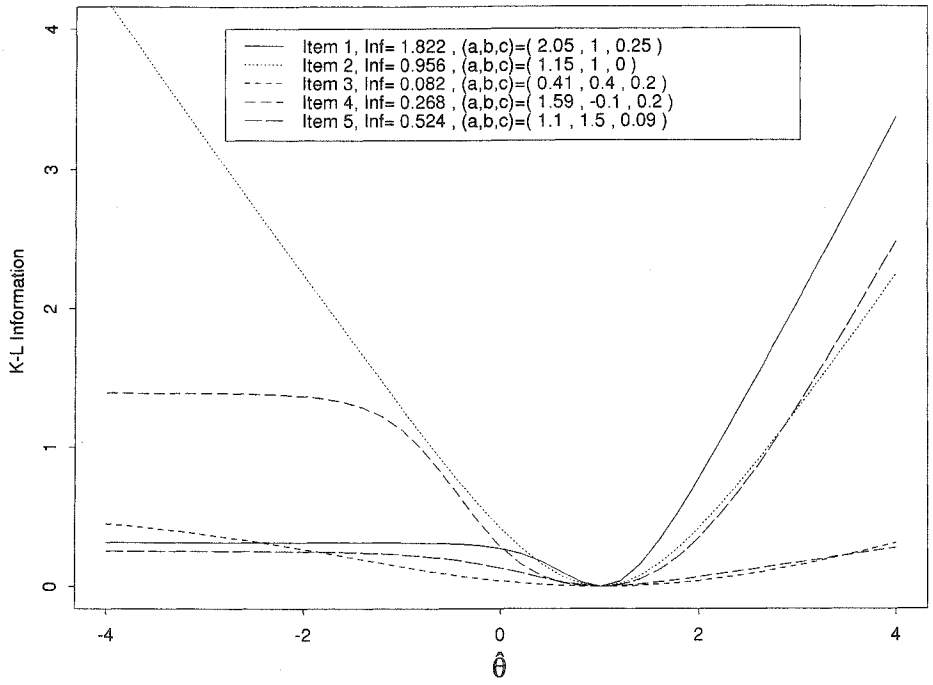
Finally, from Equation 16, I can be fully recovered from K by taking derivatives. In other words, if the profile of K is known, then I is known exactly. However, K cannot not be recovered from I. In this sense, it can be said that test or item KL information is more informative than conventional test or item Fisher information. However, K is also more complicated and, therefore, not directly applicable for obtaining a selection procedure for CAT. The main complication arises from the fact that, even with a given θ_0, K is a function on the parameter space whereas I produces a single number. Replacing θ_0 with the current estimator, the item information readily becomes an index. Hence, the next logical step is to use the KL information function to construct a summary quantity as an index.

New Item Selection Procedures for CAT

Information Index

A simple way to construct a single index from K is by taking the average over an appropriate interval of $\hat{\theta}$. An average KL information index can be defined as

Figure 2
KL Information Functions of Five Items at $\theta_0 = 1$



$$K_i(\hat{\theta}_n) = \int_{\hat{\theta}_n - \delta_n}^{\hat{\theta}_n + \delta_n} K_i(\theta \| \hat{\theta}_n) d\theta. \tag{17}$$

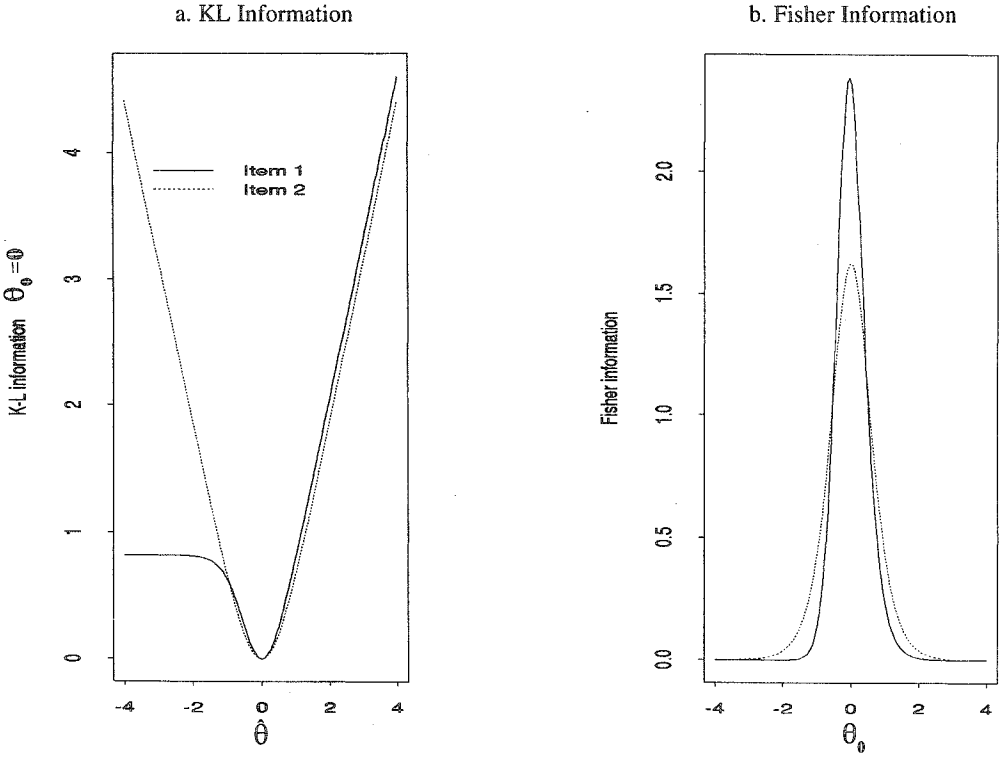
Here δ_n determines the size of the interval over which the average is computed.

The index given by Equation 17 is the area under the KL function from $\hat{\theta}_n - \delta_n$ to $\hat{\theta}_n + \delta_n$. The effect of the curvature at $\hat{\theta}_n$ is clear. For small δ_n, it is essentially determined by the curvature of $K_i(\theta \| \hat{\theta}_n)$ at $\hat{\theta}_n$. It follows that the maximum area is equivalent to the maximum curvature and hence the maximum value of I. The effect of the tails is also clear. For large δ_n, the area is also very much influenced by the tails of $K_i(\theta \| \hat{\theta}_n)$. In this respect, selection of an item based on the maximum area defined in Equation 17 reflects the idea of the global information approach.

An example showing the difference between the two item selection procedures at early stages is provided in Figure 3. Suppose both methods start with the same estimator, say $\hat{\theta} = 0$. Then, according to Figure 3b, the Fisher information method will clearly select Item 1 as the next item, because its information is larger at 0. However, the KL method (Figure 3a) will likely select Item 2 because the area under the KL information function of Item 2 becomes larger (if δ is not too small). For I, note that in this example both items reach their maximum ("informax") at 0 (−.05 for Item 2, actually). Without assuming informax, more complicated scenarios may arise (see Figure 2). Further research to gain insight into the general cases is certainly of interest.

Implementation of the average KL information index requires specifying δ_n. The preceding discussion indicates that in order to make efficient use of KL information in the context of CAT, it is reasonable to require that δ_n decrease to 0 as n approaches ∞. To determine how fast the δ_n should go to 0, recall that one of the concerns with Fisher item information is that $\hat{\theta}$ may deviate substantially from θ_0. In selecting δ_n, it is expected that the resulting interval $(\hat{\theta}_n - \delta_n, \hat{\theta}_n + \delta_n)$ will contain θ_0. It follows from general asymptotic theory for ML estimators that $\hat{\theta}_n$ is asymptotically normal with mean θ_0 and variance $1/I^{(n)}(\theta_0)$. This entails that confi-

Figure 3
Information Functions for Two Items (Item 1: $a = 2.0$, $b = -.1$, $c = .1$; Item 2: $a = 1.5$, $b = 0.0$, $c = 0.0$)



dence intervals for θ_0 should be of the type

$$\left\{\hat{\theta}_n - c\Big/\left[I^{(n)}(\hat{\theta}_n)\right]^{1/2},\quad \hat{\theta}_n + c\Big/\left[I^{(n)}(\hat{\theta}_n)\right]^{1/2}\right\} \tag{18}$$

with constant c selected according to a specified coverage probability. Because $I^{(n)}$ is of order n, it is concluded that a reasonable class for δ_n is

$$\delta_n = c/\sqrt{n}. \tag{19}$$

Note that the integration in Equation 17 is with respect to the Lebesgue measure (Billingsley, 1986) on $(\hat{\theta}_n - \delta_n, \hat{\theta}_n + \delta_n)$. The density function (up to a normalizing constant) is uniform; that is,

$$p(\theta) = 1, \quad \theta \in \left(\hat{\theta}_n - \delta_n, \hat{\theta}_n + \delta_n\right). \tag{20}$$

The Lebesgue measure was selected for convenience; other measures may also be considered.

In general, let μ_n be any probability measure on the parameter space. The associated KL index is defined as

$$K_i^{\mu_n}(\hat{\theta}_n) = \int_{\hat{\theta}_n - \delta_n}^{\hat{\theta}_n + \delta_n} K^{(n)}\left(\theta \| \hat{\theta}_n\right) d\mu_n(\theta). \tag{21}$$

This index includes Equation 17 as a special case, with μ_n taken to be the Lebesgue measure inside the interval $(\hat{\theta}_n - \delta_n, \hat{\theta}_n + \delta_n)$ and 0 the measure outside the interval.

Bayesian Information Index

If a Bayesian approach is followed, then a Bayesian information index analogous to Equation 17 may be formed. Let $\mathbf{X}_n = (X_1, ..., X_n)$. Denote $p(\theta|\mathbf{X}_n)$ as the posterior density of the parameter, which will be denoted by Θ (capitalized to indicate that it is considered as a random variable here). Define the Bayesian index for the ith item by

$$K_i^B(\hat{\theta}_n) = \int K^{(n)}(\theta \| \hat{\theta}_n) p(\theta|\mathbf{X}_n) d\theta, \tag{22}$$

where the integration is over the θ range. In practice, it is not easy to evaluate $K_i^B(\hat{\theta}_n)$, due to the fact that it is usually prohibitively difficult to compute the posterior density, especially when n is not small. This is even more problematic in this situation because the CAT must update $p(\theta|\mathbf{X}_n)$ in real time. One way to overcome the difficulty is by approximating the posterior density. According to Chang & Stout (1993),

$$P\{\Theta \le z\hat{\sigma}_n^2 + \hat{\theta}_n | X_1, \ldots, X_n\} \to \Phi(z) \tag{23}$$

as n approaches ∞, where $\hat{\sigma}_n^2 = 1/I^{(n)}(\hat{\theta}_n)$, and $\Phi(\phi)$ is the standard normal distribution (density) function. Consequently, $p(\theta|\mathbf{X}_n)$ is approximated by $\phi[(\theta - \hat{\theta}_n)/\hat{\sigma}_n]$, and an approximation to the Bayesian index $K_i^B(\hat{\theta}_n)$ can be written as

$$K_i^{AB}(\hat{\theta}_n) = \int K^{(n)}(\theta \| \hat{\theta}_n) \phi\left(\frac{\theta - \hat{\theta}_n}{\hat{\sigma}_n}\right) d\theta. \tag{24}$$

Simulation Studies

Two simulation studies were conducted to compare the global information method with the Fisher item information method. All data were generated from the 3PLM. In Study 1, the values of the item parameters were simulated from prespecified uniform distributions; in Study 2, these values were taken from a calibration of 254 items from the 1992 National Assessment of Educational Progress (NAEP) reading assessment (Johnson & Carlson, 1994).

Study 1

Item pool structure. There were 800 items in the pool. The values of the item parameters were generated from uniform distributions U(.5, 2.5), U(−3.6, 3.6), and U(0.0, .25) for a_i, b_i, and c_i, respectively. These distributions cover wide ranges of reasonable item parameters.

Test length and termination rule. Maximum test length was set at 14 items for all cases; thus, each test was terminated after the 14th item was administered. The relatively short test length was selected because interest was mainly in the performance of the item selection procedure during the early stage of CATs.

Simulation procedure. Eight different values of θ_0 were used in the simulation: $\theta_0 = -3.0, -2.0, -1.5, -1.0, 0.0, 1.0, 2.0$, and 3.0. 1,000 replications were used. The resulting ML estimators of θ_0 were denoted by $\hat{\theta}_{i,F}$ and $\hat{\theta}_{i,K}$, where subscript i indicates that the ML estimator was calculated from $(x_1, ..., x_i)$, and the subscripts F and K indicate use of the Fisher information criterion and the KL information criterion, respectively.

Initialization. For both methods, the initial item was selected with parameters $(a_1, b_1, c_1) = (a_0, b_0, c_0)$. If the outcome of the first item, X_1, was 1, then the next k_0 items were selected with increasing difficulty parameters $(b_1 \le) b_2 \le ... \le b_{k0}$ $(b_{i+1} = b_{i+2}, i \le k_0)$, where $k_0 = \min\{i: X_i = 0\}$ was the first time a 0 occurred. If the first response was a 0, then $(b_1 \ge) b_1 \ge ... \ge b_{k0}$ $(b_{i+1} = b_i - 2, i \le k_0)$ was selected, where k_0 was the first time a 1 occurred. In Study 1, $a_0 = 1$, $b_0 = -6$, and $c_0 = .2$. The as and cs remained unchanged during the initialization. Note that instead of $b_0 = 0$, the starting value for the b parameter was $b_0 = -6$, because $\theta_0 = 0$ was included in the simulation study. As a result, the CAT started with a very easy item for all eight conditions.

θ Estimation. As soon as the score sequence contained both a 0 and a 1, the ML estimator of θ was calculated. For both Fisher information-based and KL information-based selection methods, examinees' θs were estimated recursively using ML estimation (see Equations 5 or 6). More specifically, for each i, if the components in $(x_1, ..., x_i)$ were not all the same, the ML estimator $\hat{\theta}_i$ was calculated according to a numerical algorithm that mimicked a subroutine in the LOGIST program (Wingersky, Barton, & Lord, 1982).

The ML estimation algorithm used standard Newton-Raphson iterations (Cheney & Kincaid, 1985). When the 3PLM is used, it may have multiple roots (Samejima, 1973). Thus, instability may be encountered at early stages of the estimation. However, no multiple-roots searching technology was used here. Further discussions concerning improving the stability of ML estimation calculation can be found in Chang & Ying (1996a).

Item selection. Given a ML estimate of θ, for the Fisher information-based method the $(i + 1)$th item was selected such that $I_{i+1}(\hat{\theta}_i)$ had the maximum value among all items in the pool; for the KL information-based method, the $(i+1)$th item was selected such that $K_{i+1}(\hat{\theta}_i)$ had the maximum value. Each time an item was used, it was then deleted from the item pool. For Study 1, $\delta_n = 3/(n)^{1/2}$, in accordance with Equation 18, and $c = 3$.

Evaluation criteria. $\hat{\theta}_{i,F}$ and $\hat{\theta}_{i,K}$ were calculated and their bias functions were calculated by

$$\mathrm{BIAS}_F(i) = \frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\hat{\theta}_{i,F} - \theta_0, \quad i = 5, \ldots, 14, \tag{25}$$

and

$$\mathrm{BIAS}_K(i) = \frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\hat{\theta}_{i,K} - \theta_0, \quad i = 5, \ldots, 14. \tag{26}$$

The mean squared errors (MSEs) also were calculated for every i, $i = 5, ..., 14$. The MSEs were defined by

$$\mathrm{MSE}_F(i) = \frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\left(\hat{\theta}_{i,F} - \theta_0\right)^2, \quad i = 5, \ldots, 14, \tag{27}$$

and

$$\mathrm{MSE}_K(i) = \frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\left(\hat{\theta}_{i,K} - \theta_0\right)^2, \quad i = 5, \ldots, 14. \tag{28}$$

Note that if the test length is short, ML estimation might not provide a solution. For both MSE and bias, only those $\hat{\theta}$s with $i \geq 5$ were used.

Results. Figures 4 and 5 summarize the simulation results. Under several of the eight simulation conditions, both average bias (Figure 4) and MSE (Figure 5) were uniformly smaller for item selection using KL information than using Fisher information. For example, in three of the eight cases in Figure 4 the KL method resulted in substantial bias reduction ($\theta_0 = -3, -2, -1.5$), while in the remaining cases the performance of KL was either slightly better or similar to that of Fisher. Improvements in terms of MSEs was either more significant or similar, as shown in Figure 5.

Study 2

The methods used in Study 2 were essentially the same as those in Study 1. The differences were (1) the θ_0 range was −2.0, −1.0, 1.0, and 2.0; (2) the starting value for the b parameter was $b_0 = 0$; (3) the test length was set to $n = 40$; and (4) the item parameters were taken from the Reading Assessment of the 1992 NAEP main assessment sample (Johnson & Carlson, 1994). For the 254 items, 122 had parameter estimates from the two-parameter logistic model, and 132 had parameter estimates from the 3PLM. These parameters were not uniformly distributed, as can be seen from the histograms of the parameter distributions (Figure 6).

Figures 7 and 8 summarize the results of Study 2. In two of the four cases summarized in Figure 7, KL gave better bias reduction ($\theta_0 = -2, -1$). There was essentially no difference between the two methods for the

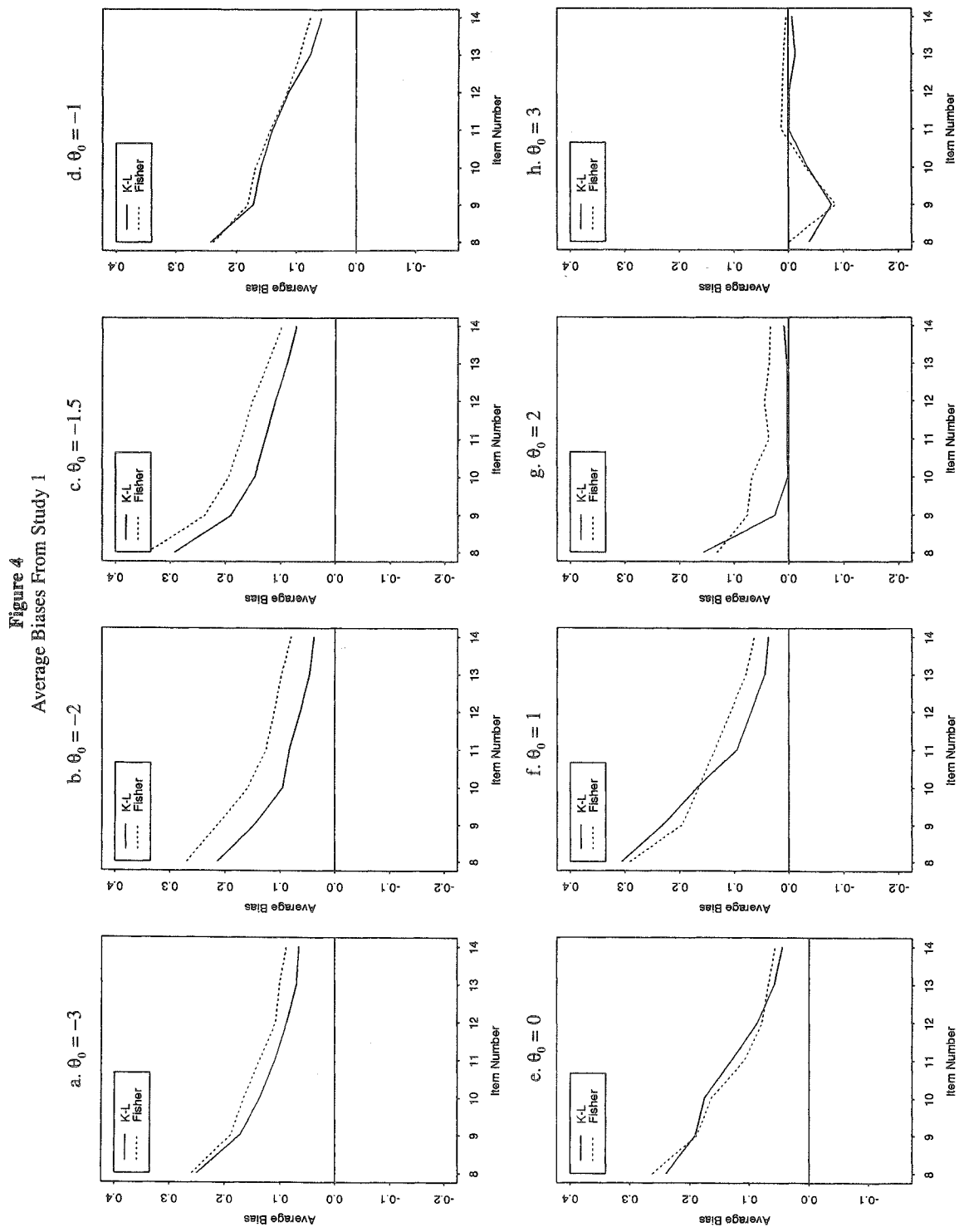


Figure 4
Average Biases From Study 1

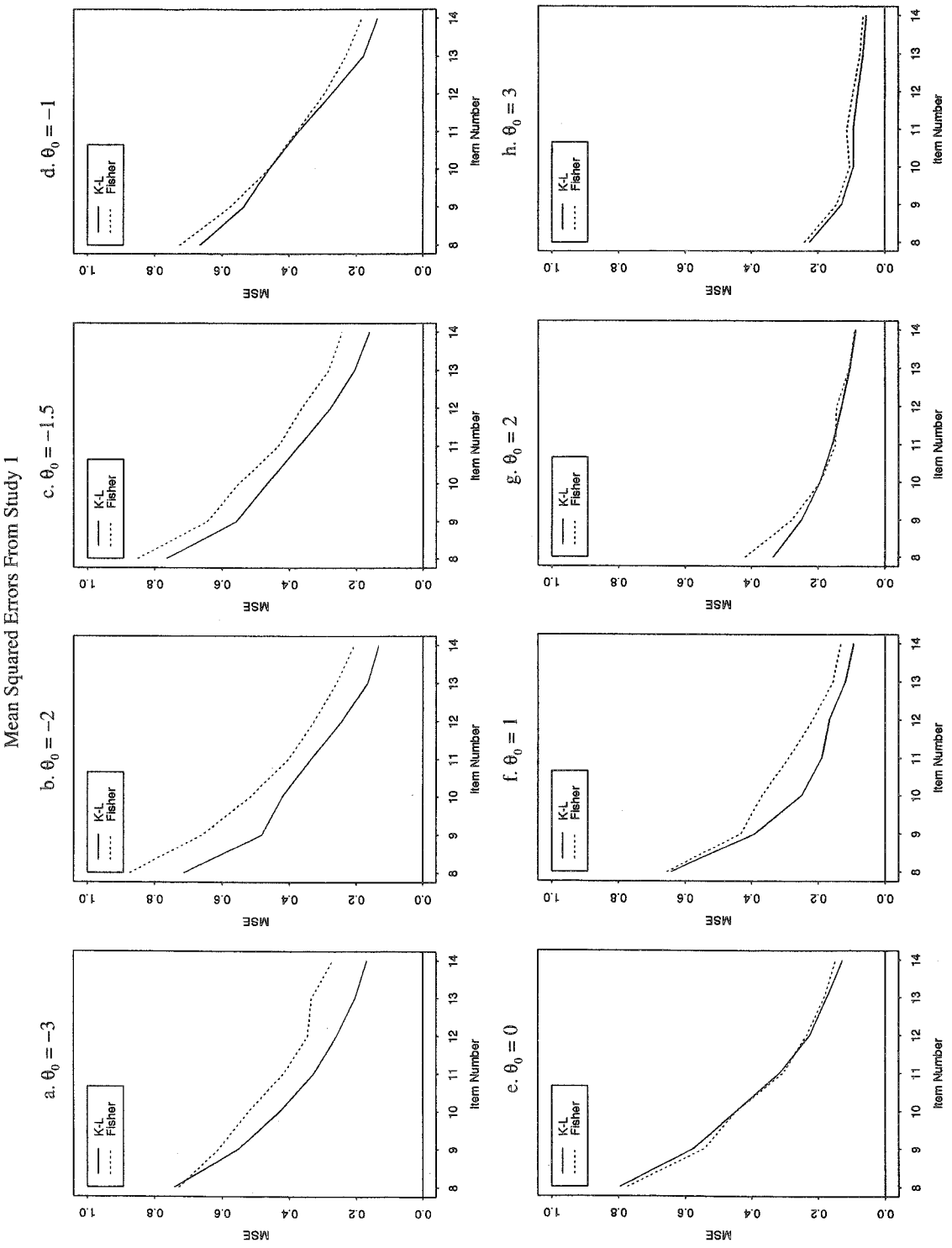
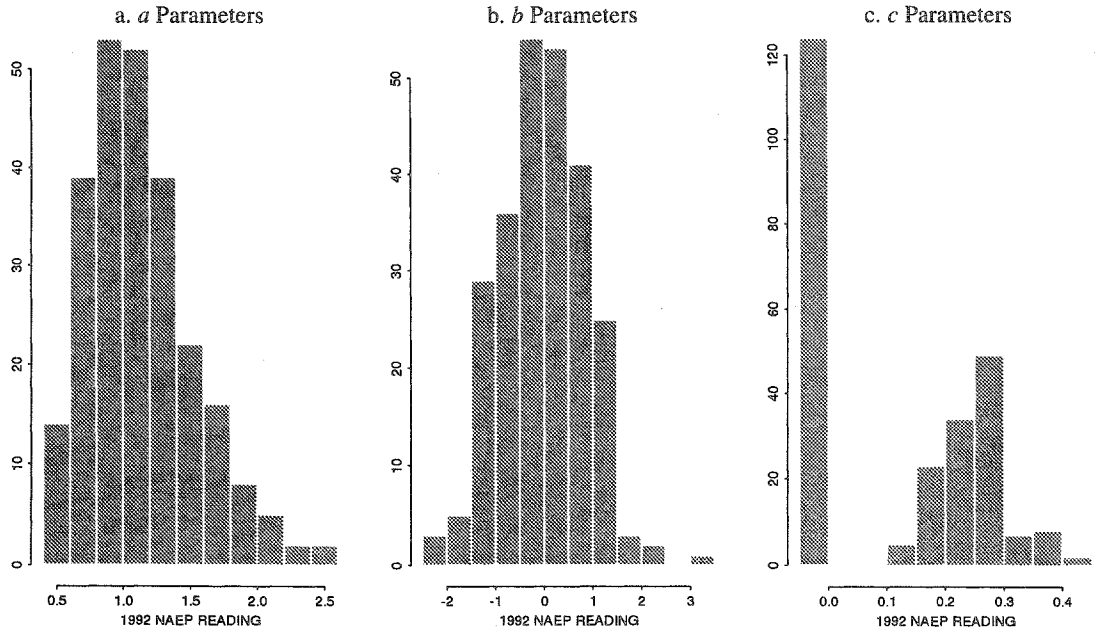


Figure 5
Mean Squared Errors From Study 1

Figure 6
Histograms of the Item Parameter Distributions in Study 2



remaining two cases. Figure 8 indicates that the reduction of MSEs was significant only for one of the four cases ($\theta_0 = -2$). For the remaining three cases, the reduction was pronounced only when n was small (i.e., in the early stages of a CAT). For $n > 30$, there was essentially no difference. This may be expected because, for large n, KL should be equivalent to that of Fisher.

Discussion

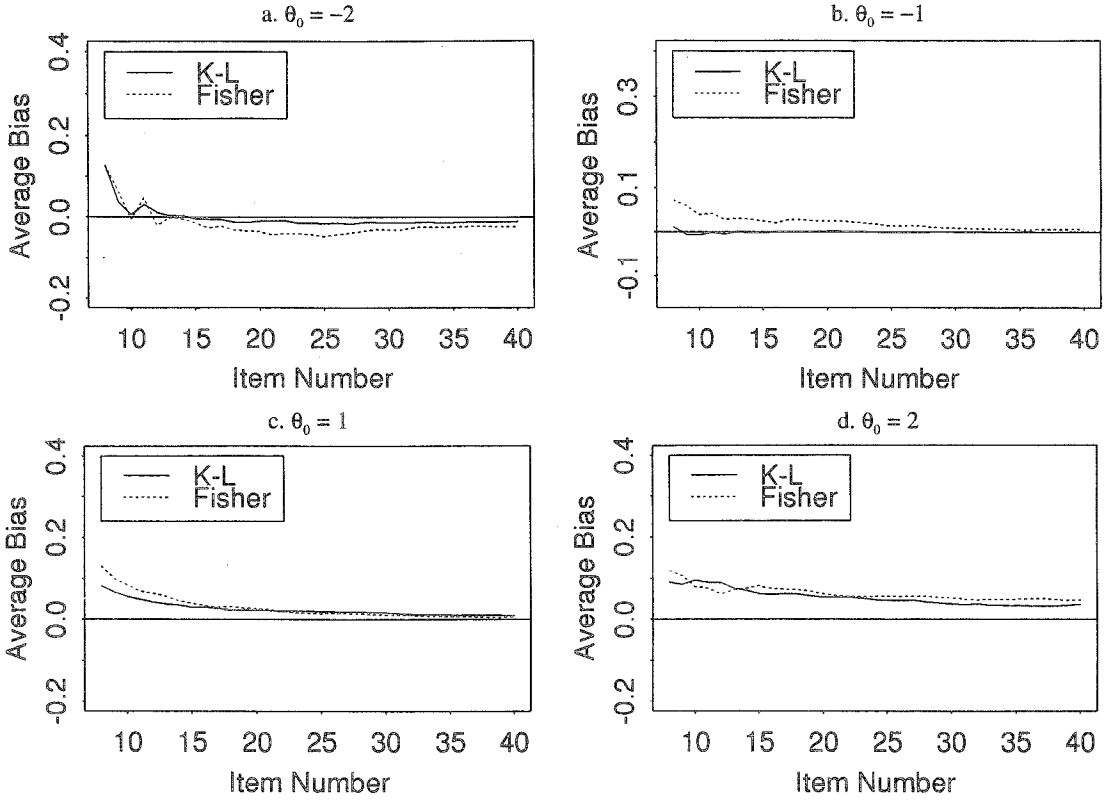
The results of the pilot simulation studies indicated that the proposed global information index is a promising alternative to the Fisher information-based item selection methods. The performance of the KL approach was slightly better than that of Fisher. In many cases, the KL approach outperformed the Fisher method in terms of bias reduction and smaller mean squared errors. The improvements were rather noticeable at early stages of the simulated tests.

Many issues remain to be investigated. Both the global information and Fisher information selection procedures lack theoretical justification. The main difficulty results from the dependent structure arising from sequential item selection. Recently, however, Chang & Ying (1996a, 1996b) demonstrated that recursively calculated ML estimators are consistent and asymptotically normal under suitable regularity conditions.

To explore the full capacity of KL information, more extensive simulation studies are needed. The choice for the bandwidth δ_n deserves special attention. Moreover, because K is not symmetric about θ_0, it is reasonable to consider nonsymmetric averaging in Equation 17 (i.e., integrate from $\hat{\theta}_n - \delta_{n,1}$ to $\hat{\theta}_n + \delta_{n,2}$ with $\delta_{n,1} \neq \delta_{n,2}$).

Both local and global information can be projected together into a three-dimensional space. Because the true parameter θ_0 is unknown, it may be more informative to consider KL information as a function of two variables. In other words, consider $K(\hat{\theta}\|\theta_0)$ as a function of both $\hat{\theta}$ and θ_0. This effectively creates a surface in three-dimensional Euclidean space, where the third axis is the value of $K(\hat{\theta}\|\theta_0)$. The geometry of this is as follows: KL information functions for different θ levels are the slices of the information surface. For

Figure 7
Average Biases From Study 2



example, functions $K(\cdot\|\theta_0)$ and $K(\cdot\|\theta_0')$ with θ_0 and θ_0' fixed are the KL information functions for θ_0 and θ_0', respectively. Note that the curvature at $\theta_0 = \hat{\theta}$ of the function intersected by the surface and the vertical plane $\lambda = \theta_0$ is Fisher information at θ_0 (see Figure 1). In this connection, another new index can be defined that represents the volume under the information surface:

$$\overline{K}(\hat{\theta}_n) = \int_{\hat{\theta}_n-\eta_n}^{\hat{\theta}_n+\eta_n}\int_{\hat{\theta}_n-\delta_n}^{\hat{\theta}_n+\delta_n} K(\theta_1\|\theta_2)d\hat{\theta}_1 d\theta_2, \tag{29}$$

where η is a quantity similar to δ_n, but may be independent of δ_n. Note that the uniform density can be replaced by a general measure.

Finally, this conceptualization of global information may change the traditional view of low discriminating items. Figure 9 indicates that if there is little knowledge about the location of θ_0, then an item with a low *a* parameter (Figure 9b) may be a better choice for the examinee than an item with a high *a* parameter (Figure 9a). Note that for any $\theta_0 \neq \hat{\theta}$, the item in Figure 9b tends to contain a certain amount of global information and, thus, is more robust. However, the item in Figure 9a has adequate information content only in part of the region. It delivers almost no information for approximately 50% of the entire region. However, if the specific range around θ_0 is known, say around 0, then it will be more efficient to select the item in Figure 9a for the examinee.

Figure 8
Mean Squared Errors From Study 2

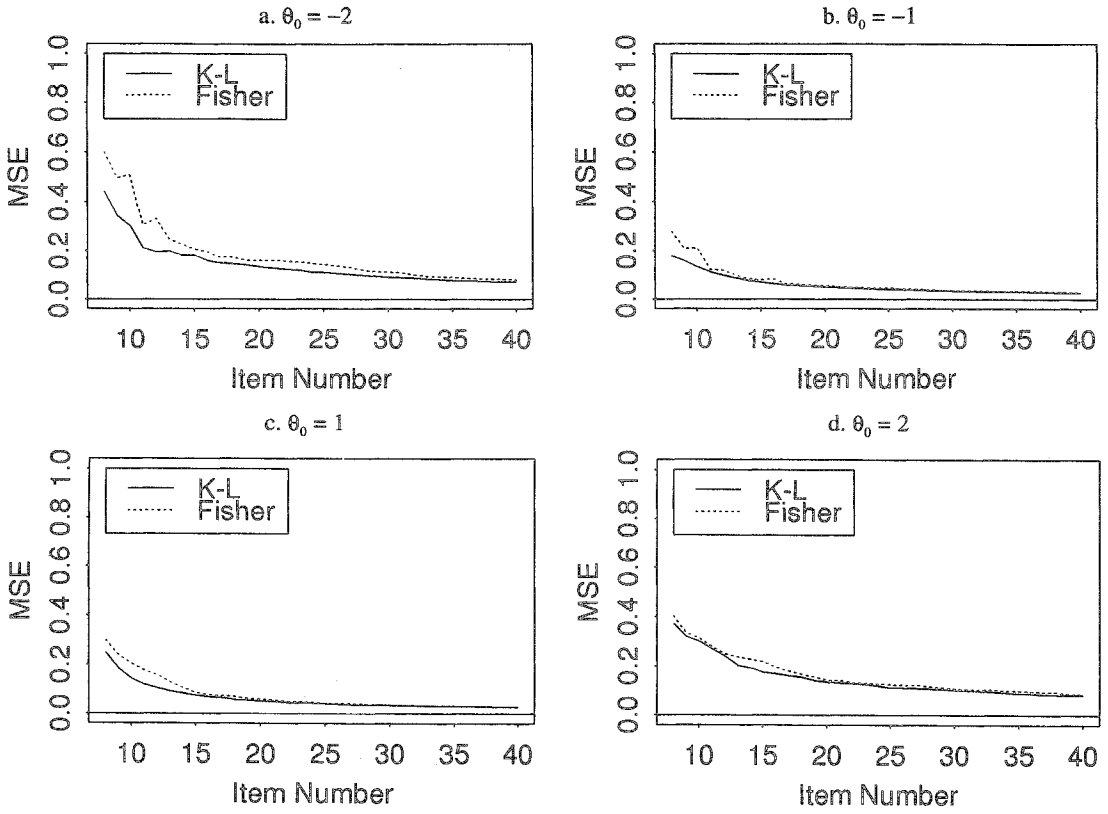
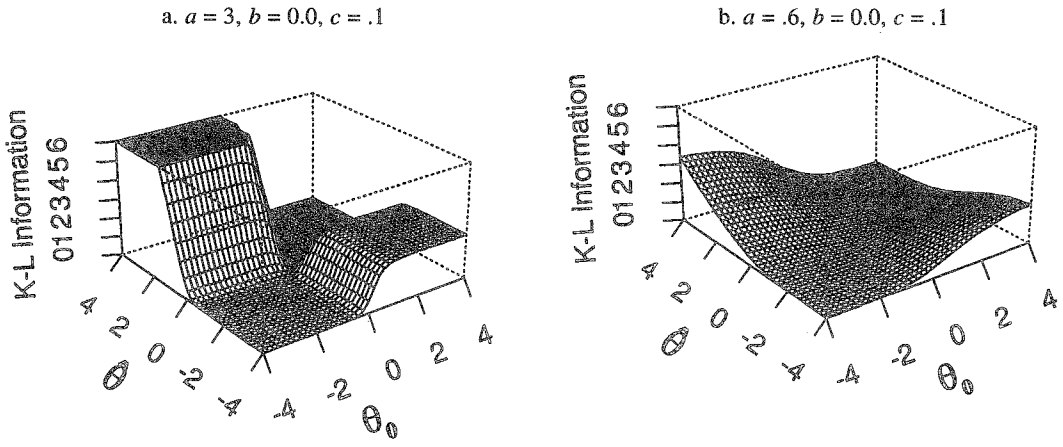


Figure 9
Global Information Surfaces for Two Items



References

Billingsley, P. (1986). *Probability and measure.* New York: Wiley.

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

Chang, H.-H., & Ying, Z. (1996a). *Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests.* Manuscript submitted for publication.

Chang, H.-H., & Ying, Z. (1996b, June). *Building a statistical foundation for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Banff, Alberta, Canada.

Cheney, W., & Kincaid, D. (1985). *Numerical mathematics and computing.* Monterey CA: Brooks/Cole.

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

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Fan, M., & Hsu, Y. (1995, June). *The effect of ability estimation for polytomous CAT in different item selection procedures.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis MN.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff.

Johnson, E. G., & Carlson, J. E. (1994). *The NAEP 1992 technical report.* Washington DC: National Center of Education Statistics.

Kullback, S. (1959). *Information theory and statistics.* New York: Wiley.

Le Cam, L., & Yang, G. L. (1990). *Asymptotics in statistics: Some basic concepts.* New York: Springer-Verlag.

Lehmann, E. L. (1983). *Theory of point estimation.* New York: Wiley.

Lehmann, E. L. (1986). *Testing statistical hypotheses.* New York: Wiley.

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

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison Wesley.

Neyman, J., & Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. I. Unbiased critical regions of type A and type A_1. *Statistical Research Memorandum, 1,* 1–37.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38,* 221–233.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* New York: Wiley.

Stocking, M. L. (1993, February). *Modern computerized adaptive testing.* Paper presented at the Joint Statistics and Psychometrics Seminar, Princeton NJ.

van der Linden, W. J. (1995, June). *Bayesian item selection in adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis MN.

Veerkamp, W. J., & Berger, M. P. F. (1994). *Some new item selection criteria for adaptive testing* (Research Rep. 94-6). Enschede, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale NJ: Erlbaum.

Weiss, D. J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (pp. 24–35). Washington DC: United States Civil Service Commission.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide* [Computer program manual]. Princeton NJ: Educational Testing Service.

Acknowledgements

This research was partially supported by Educational Testing Service Allocation Project No. 79427, and the National Assessment of Educationl Progress (Grant No. R999J40001 and CFDA No. 84.999J) as administered by the Office of Educational Research and Improvement, U.S. Department of Education, by the National Science Foundation, and by the National Security Agency. The authors thank Erich Lehmann, Barbara Dodd, Bert Green, Xuming He, Frank Jenkins, Charles Lewis, Spence Swinton, Howard Wainter, Bo Wang, and Ming-Mei Wang for many helpful comments and discussions. They particularly thank the Editor and two anonymous reviewers for their suggestions, which led to numerous improvements.

Author's Address

Send requests for reprints or further information to Hua-Hua Chang, Mail Stop 02-T, Educational Testing Service, Princeton NJ 08541, U.S.A., or to Zhiliang Ying, Department of Statistics, Rutgers University, Hill Center, Busch Campus, New Brunswick NJ 08903, U. S. A. Email: hchang@ets.org. or zying@stat.rutgers.edu.