# A Use of the Information Function in Tailored Testing

**Fumiko Samejima**
**University of Tennessee**

Several important and useful implications in latent trait theory, with direct implications for individualized adaptive or tailored testing, are pointed out. A way of using the information function in tailored testing in connection with the standard error of estimation of the ability level using maximum likelihood estimation is suggested. It is emphasized that the standard error of estimation should be considered as the major index of dependability, as opposed to the reliability of a test. The concept of weak parallel forms is expanded to testing procedures in which different sets of items are presented to different examinees. Examples are given.

Researchers have tended to use latent trait theory rather than classical test theory in research on individualized adaptive or tailored testing. This is quite natural, since latent trait theory has definite merits over classical test theory in many crucial matters.

Because of the lack of opportunities to really learn the theory, however, these researchers tend to overlook some important implications in latent trait theory. As a result, its full use has not yet materialized. Not only are information functions seldom used to maximum advantage, but also those who have tried to use latent trait theory still use some popular concepts in classical test theory, such as reliability.

In this paper, the author points out some important implications in latent trait theory which are not fully understood and appreciated among researchers, and gives some practical suggestions for its use.

## Reliability and Standard Error of Estimation in Classical Test Theory

The reliability of a test has been considered as one of the central concepts in classical test theory, and much research has been conducted on this and related concepts. The measure of the reliability of a test is usually a simple correlation coefficient between the two sets of test scores obtained by two parallel measurements, or its estimated value using the Spearman-Brown formula. Since it is a correlation coefficient, it depends not only upon the test itself but also upon the specific group of examinees tested. This fact has often been overlooked, and reliability has been treated as if it were an intrinsic parameter of the test. To give an extreme example, however refined the test may be, the reliability coefficient is zero if all examinees have exactly the same true score. Conversely, it is easy to make a poorly constructed test look good by calculating the correlation coefficient for a group of examinees whose ability levels are substantially different from one another. A crucial shortcoming of clas-

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 1, No. 2 Spring 1977 pp. 233–247*
*© Copyright 1977 West Publishing Co.*

sical test theory is that all the important indices, including reliability, coefficient alpha, and Kuder-Richardson formulas 20 and 21, are defined within a specific group of examinees; therefore, their generalizability is limited. Yet people tend to attach absolute meanings to these concepts. This is a non-scientific and undesirable tendency.

There have been several different indices of the error of estimation in classical test theory. For instance, in Lord and Novick (1968, Ch. 3), the standard error of measurement $\sigma_E$ is given by

$$\sigma_E = \sigma_X[1 - \rho_{XX'}]^{1/2} \quad , \qquad [1]$$

where $\sigma_X$ is the standard deviation of the observed test score $X$, and $\varrho_{xx'}$ is the reliability coefficient of the test, the equation which is most widely used. The standard error of estimation of the true score by linear regression estimation is given by

$$\sigma_\varepsilon = \sigma_X[\rho_{XX'}(1 - \rho_{XX'})]^{1/2} \quad , \qquad [2]$$

which has a slightly different meaning.

In spite of the differences in these indices, one thing common to all is that they depend on a particular group of examinees, as the reliability coefficient does, and in most cases they are derived from the reliability coefficient. Thus their generalizability is again limited, and we cannot treat them as intrinsic parameters of a given test. In addition, since these indices are given for the whole group of examinees, the best we can do is *assume* the independence of true scores and error scores and use the same index conditionally, too. Even so, this assumption of independence is hardly acceptable for true scores and error scores (see Samejima, 1976a).

The situation is substantially different in latent trait theory where the standard error of estimation does have an intrinsic meaning, since test information function is defined independently of any specific group of examinees. Before going into detail, however, a brief review of information functions may be helpful.

## Information Functions

One of the important and useful concepts in latent trait theory is a group of information functions. The *item information function* for a binary item was defined by Birnbaum (1968, p. 454) as

$$I(\theta, u_g) = P'_g(\theta)^2[P_g(\theta)\ Q_g(\theta)]^{-1} \quad , \qquad [3]$$

where $\Theta$ is the unidimensional ability to be measured, $u_g$ is the binary item score of item $g$, $P_g(\Theta)$ is the item characteristic function of item $g$, $P'_g(\Theta)$ is its first derivative with respect to $\Theta$, and $Q_g(\Theta)$ is $1 - P_g(\Theta)$. Birnbaum also defined the *test information function* of a test consisting of $n$ binary items as

$$I(\theta) = \sum_{g=1}^{n} I(\theta, u_g) \quad . \qquad [4]$$

Samejima (1969, Ch. 6) defined the *item response information function* $I_{x_g}(\Theta)$ for a graded item score category $x_g$ of item $g$ as

$$\begin{aligned} I_{x_g}(\theta) &= -\frac{\partial}{\partial\theta} A_{x_g}(\theta) \\ &= -\frac{\partial^2}{\partial\theta^2} \log P_{x_g}(\theta) \quad , \qquad [5] \end{aligned}$$

where $P_{x_g}(\Theta)$ is the operating characteristic of the graded score category $x_g$, which assumes an integer 0 through $m_g$ according to the degree of attainment to the solution of item $g$, and $A_{x_g}(\Theta)$ is the basic function which is the ratio of the first derivative of $P_{x_g}(\Theta)$ with respect to $\Theta$ to $P_{x_g}(\Theta)$ itself. When the item response is a continuous variable $z_g$, the item response information function is given by

$$\begin{aligned} I_{z_g}(\theta) &= -\frac{\partial}{\partial\theta} A_{z_g}(\theta) \\ &= -\frac{\partial^2}{\partial\theta^2} \log H_{z_g}(\theta) \quad , \qquad [6] \end{aligned}$$

where $H_{z_g}(\Theta)$ is the operating density characteristic of item response $z_g$ and $A_{z_g}(\Theta)$ is the basic function of $z_g$ (Samejima, 1973a). The item information function for a graded response item is given by

$$I_g(\theta) = E[I_{x_g}(\theta)]$$

$$= \sum_{x_g=0}^{m_g} I_{x_g}(\theta)\, P_{x_g}(\theta) \quad , \qquad [7]$$

and it has been pointed out that Birnbaum's item information function, which is given by Equation 3, is a special case of Equation 7 when $m_g = 1$ (Samejima, 1969, pp. 39-40). The item information function for a continuous response item can be written as

$$I_g(\theta) = E[I_{z_g}(\theta)] \qquad [8]$$

$$= \int_0^1 [-\frac{\partial^2}{\partial\theta^2} \log H_{z_g}(\theta)] H_{z_g}(\theta)\, dz_g \ .$$

The test information function for a test consisting of $n$ graded items is given in relation to the likelihood function of the response pattern $V$ by

$$I(\theta) = E[\frac{\partial}{\partial\theta} \log L_V(\theta)]^2 \qquad [9]$$

$$= - E[\frac{\partial^2}{\partial\theta^2} \log L_V(\theta)] = \sum_{g=1}^{n} I_g(\theta).$$

Similarly, for a test of $n$ continuous items we have

$$I(\theta) = \sum_{g=1}^{n} I_{z_g}(\theta) \quad . \qquad [10]$$

There are many useful implications in the information functions. For instance, the item response information function gives a measure of steepness of the basic function, which is a component of the derivative of the log likelihood in the likelihood equation such that

$$\frac{\partial}{\partial\theta} \log L_V(\theta) = \sum_{x_g \varepsilon V} A_{x_g}(\theta) = 0. \qquad [11]$$

It has also been pointed out that the basic function on the continuous response level is the *asymptotic basic function* (Samejima, 1972) on the graded response level, and if this function is strictly decreasing in Θ, then the unique local maximum for the likelihood function is assured for every response pattern on the continuous response level, as well as for every response pattern (except for the two extreme cases where all the item scores are zero or the full item scores) on the graded response level. In such a case, the item response information function is positive almost everywhere, and this condition is satisfied in both the normal ogive and the logistic models, but not in the three-parameter normal or logistic model. In the latter case the unique maximum is not assured for every response pattern (Samejima, 1973b). This fact deserves special emphasis, since, without considering the item response information function, we may be tempted to use, for instance, the item information function I(Θ, $u_g$) as the measure of accuracy in maximum likelihood estimation of the examinee's ability, and from Equation 3 it is obvious that I(Θ, $u_g$) is always non-negative, whether both of the item response information functions are positive or not.

The test information function provides the limit of the amount of information given by a test in which any scoring strategy is used, as Birnbaum has pointed out (Birnbaum, 1968, p. 454). Perhaps the most useful property of the test information function is, however, that the conditional distribution of the maximum likelihood estimate $\hat{\Theta}$ is *asymptotically* normal with

$$E(\hat{\theta}|\theta) \simeq \theta \qquad [12]$$

and

$$Var.\ (\hat{\theta}|\theta) \simeq [I(\theta)]^{-1} \ , \qquad [13]$$

when the number of items increases (Birnbaum, 1968, p. 457; Samejima, 1975; cf. Cramer, 1946, Section 33.3). In fact, this is the property which can be effectively used in the tailored testing situation, as we shall see in the following sections.

## Standard Error of Estimation in Latent Trait Theory

In contrast to classical test theory, latent trait theory gives us the standard error of estimation free from any specific group or population of examinees. Let $\varepsilon$ be the discrepancy between the maximum likelihood estimate $\hat{\Theta}$ and ability $\Theta$, such that

$$\hat{\theta} = \theta + \varepsilon \quad . \qquad [14]$$

By virtue of the asymptotic property of the maximum likelihood estimate described at the end of the preceding section, the distribution of the error $\varepsilon$ is asymptotically $N(O, I(\Theta^{-1})$, and $I(\Theta)^{-1/2}$ can be used as the standard error of estimation when the number of items and the amount of information given by the test are sufficiently large. This function does not depend on any specific group of examinees, as is obvious from Equations 4, 10 and 11.

It might be asked: How large should $I(\Theta)$ be in order to tolerate the approximate normal distribution for the error $\varepsilon$ ? To answer this, we must depend, more or less, on our subjective judgment. Some examples were given elsewhere (Samejima, 1975, 1976a) in which 100 maximum likelihood estimates were produced on the same ability level by the monte carlo method and the cumulative frequency distributions were given together with the normal distribution functions. In these examples, the values of the test information functions were approximately 22.1, 17.3 and 7.4 respectively. Figure 1 shows two more such examples, in which $I(\Theta) \doteq 21.8$ at $\Theta = -0.6$ and $I(\Theta) \doteq 12.4$ at $\Theta = 1.5$ respectively. We can see that in these examples the fits are fairly good.

Another important point is whether or not $E(\hat{\Theta}|\Theta)$ has a systematic bias, or $E(\varepsilon|\Theta)$ is a function of $\Theta$. If this is the case, then even if Equations 12 and 13 are asymptotically true, the approximate characteristics of the behavior of $\varepsilon$, and hence that of $\hat{\Theta}$, which will be presented and discussed in later sections, will not necessarily hold. However, observation of the five maximum likelihood estimates, which were produced by the monte carlo method for each of one hundred ability levels equally spaced between $\Theta = -2.475$ and $\Theta = 2.475$, contradicts such a possibility, and from this result there is no reason to believe that $E(\varepsilon|\Theta)$ varies as a function of $\Theta$ when $I(\Theta)$ is as large as 21.6 (Samejima, 1976c).

Three similar sets of results, which will be given in the following section of this paper (Tables 2 and 3) with $I(\Theta) = 25$, 20 and 16 respectively, also contradict this possibility, and again there is no reason to believe that $E(\varepsilon|\Theta)$ varies as a function of $\Theta$ with these values of $I(\Theta)$. On the other hand, the distribution of $\varepsilon$ was tested against the null hypothesis, $N(0, I(\Theta)^{-1})$, for each of the ten different levels of ability in the above three cases, using the chi-square test with 2 degrees of freedom in the first case and 1 in the other two cases. The ten results of each case were categorized into four classes with respect to their probabilities obtained by the test, which turned out to be 3, 2, 3 and 2; 4, 3, 2 and 1; and 2, 3, 3 and 2 in frequencies for the intervals of probability, (.00, .25), (.25, .50), (.50, .75) and (.75, 1.00) respectively. Since the expected frequencies are 2.5, 2.5, 2.5 and 2.5 for all cases, these results suggest the acceptability of the null hypothesis in all cases. The chi-square test was also conducted for the total ten ability levels with 20, 10 and 10 degrees of freedom, and the resulting values of the test statistic were 19.5900, 13.6440 and 8.5632 respectively. In all three cases they were not statistically significant at the .10 level.

From these findings, we will assume that $E(\varepsilon|\Theta)$ does not vary as a function of $\Theta$. Equations 12 and 13 are both sufficiently satisfied with our data. (For simplicity, in the remaining sections of the paper, the symbol $\simeq$, which is used in Equations 12 and 13, will be replaced by an equality sign.)

Thus, unlike classical test theory, latent trait theory provides us with the standard error of estimation as a measure independent of any particular group of examinees, and given locally, or as a function of ability $\Theta$. For this reason, we can

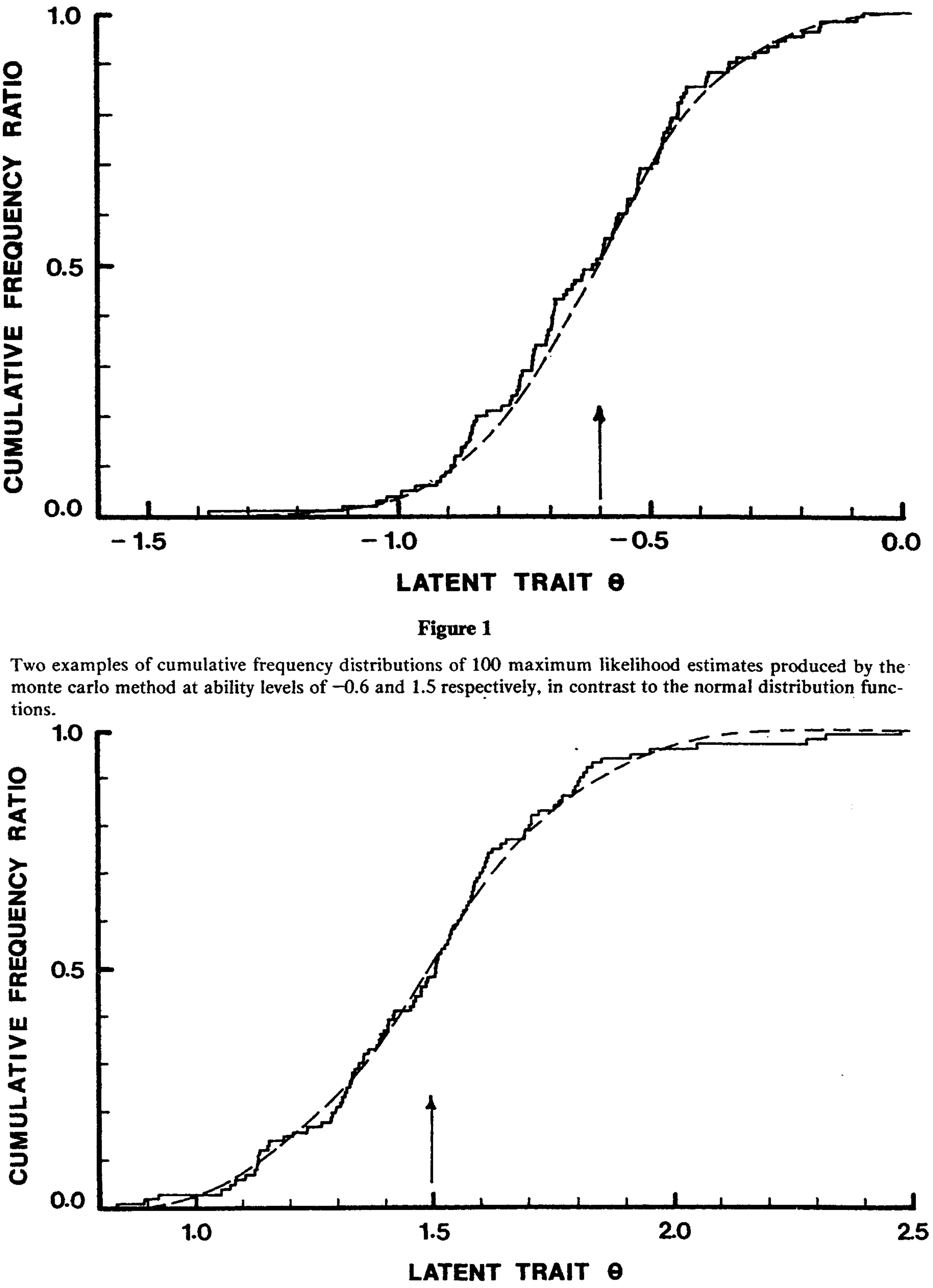


**Figure 1**

Two examples of cumulative frequency distributions of 100 maximum likelihood estimates produced by the monte carlo method at ability levels of −0.6 and 1.5 respectively, in contrast to the normal distribution functions.

consider this as an intrinsic property of the test itself, as long as the populations of our interest belong to the complete latent space (Lord and Novick, 1968, pp. 359-360).

## Weak Parallel Forms and Tailored Testing

In latent trait theory, parallel forms are defined as a pair of tests which measure the same ability, and whose items have a one-to-one correspondence with respect to their identity of the numbers of score categories and their operating characteristics. Samejima (1976a, b) has suggested the term *strong parallel forms* for this situation, and has defined *weak parallel forms* as a pair of tests which measure the same ability and whose test information functions are identical. Thus two tests can be parallel, regardless of the possible differences in the numbers of items, in their scoring strategies and operating characteristics, and in other characteristics.

This definition of weak parallel forms goes beyond the level of fixed item tests, such as paper-and-pencil tests, and is applicable to tailored testing. Thus in spite of the fact that in individualized adaptive testing no fixed set of items are given to all the examinees, as long as the test information function is kept identical for two sessions of administration, they are considered as weak parallel forms.

In computer-assisted individualized adaptive testing, one criterion for terminating the presentation of new items is when the change in the values of the maximum likelihood estimate becomes less than a certain small value. Another method, which is probably more logical and justifiable, may be to terminate the presentation when the test information function has reached a certain value or when the standard error of estimation has reached a certain small value. This critical value can be a constant for all the levels of ability, or can be defined locally as a function of ability $\Theta$.

To give an example, suppose that our purpose for testing is selection, and we want to accept any candidate whose ability level is greater than or equal to a certain critical value, and reject anyone whose ability level is less than that value. Figure 2 presents the amount of test information necessary to make such a selection with the levels of confidence .841, .933, .977, .994 and .999 respectively, using the normal approximation to the conditional distribution of the error $\varepsilon$. Figure 3 presents the corresponding standard errors of estimation for the above five cases. In these cases the critical value of $\Theta$ is 1.5. Thus as long as we use these locally determined amounts of information, or the standard errors of estimation, as our criterion in terminating testing, two or more such sessions are weakly parallel. With this kind of strategy, examinees whose ability levels are far more or far less than 1.5 do not have to try many test items, but those who are close to 1.5 in their ability have to take many test items. In fact, theoretically, anyone whose ability is exactly 1.5, must be presented infinitely many test items; in practice, some suitable compromise will have to be made. It should also be kept in mind that our observation is $\hat{\Theta}$, not $\Theta$ itself, so the termination of presenting new items should wait until the amount of information has reached the criterion value at the level of $\Theta$ which is somewhat closer to the critical value 1.5 than the examinee's current maximum likelihood estimate. This will maintain the accuracy of selection.

If we want to measure different ability levels with equal accuracies, then we should set the criterion information constant at every ability level in our consideration. In such a case, we have an additional advantage that the distributions of the error $\varepsilon$ and ability $\Theta$ are *almost* independent. This situation is beneficial when we want to estimate the operating characteristics of new items which are to be added to our item library, as was pointed out by Samejima (1976b, c). If we use such a criterion for selection purposes, it is much less likely that the candidates whose ability levels are far less than the critical value will be accepted, or that those whose ability levels are much higher than the critical value will be rejected. It will waste computer time, however, by measuring too accurately applicants' ability levels which are far from the criti-

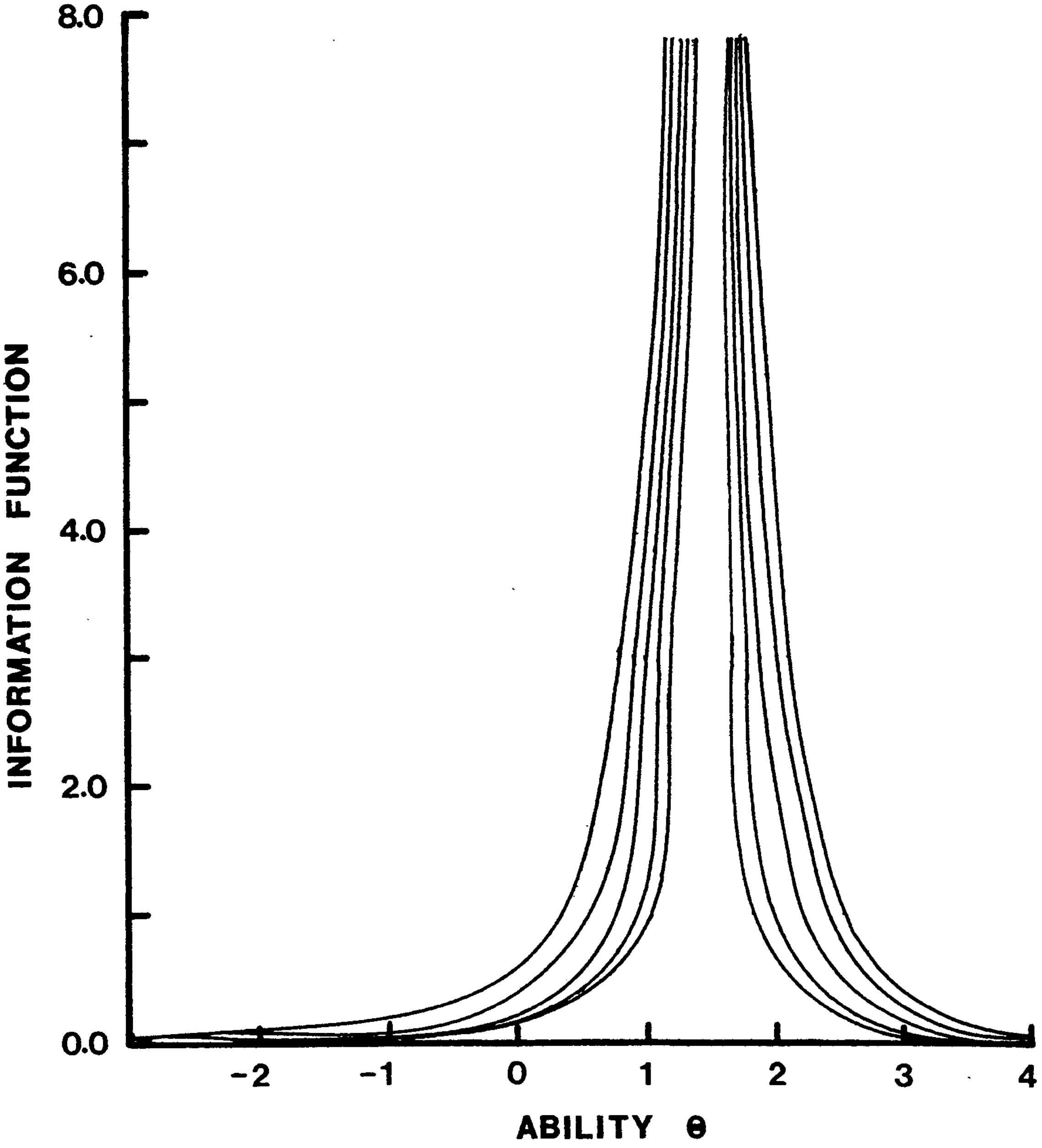


**Figure 2**

Five test information functions required to make a selection at the ability level of 1.5, with the levels of confidence 0.841, 0.933, 0.994 and 0.999 respectively.

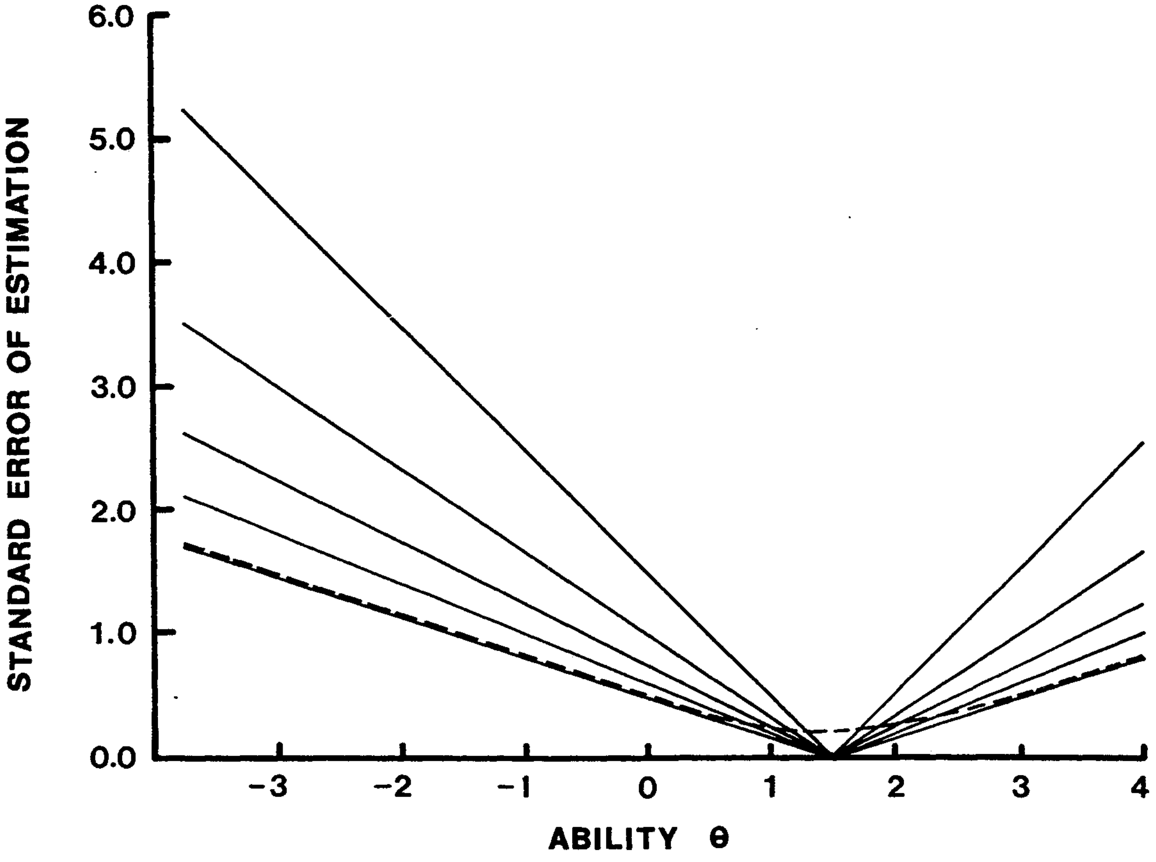


**Figure 3**

Five curves of the standard error of estimation corresponding to the five test information functions presented in Figure 1.

cal value.

On the other hand, if we use one of the strategies shown in Figures 2 and 3, it is equally probable that an applicant whose ability level is far less than the critical value will be accepted, as one whose ability level is close to the critical point, although the probability is very small. Perhaps a good compromise is to use the strategy or criterion shown by the dashed line in Figure 3. Following this curve, those whose ability levels are less than 0.0 will be accepted, and those whose ability levels are greater than 3.0 will be rejected, with the probability a little more than .0013. If we take the values of ability 1.0 and 2.0 instead of 0.0 and 3.0 respectively, the probability is no greater than .0062, and at Θ = 1.3 and Θ = 1.7, this probability is .1587. At Θ = 1.5 the chances are equal.

To give another example, consider a hypothetical tailored testing situation in which there is a set of many binary items, whose item characteristic functions follow the normal ogive model such that

$$P_g(\theta) = [2\pi]^{-1/2} \int_{-\infty}^{a_g(\theta-b_g)} \exp[-t^2/2]\, dt, \qquad [15]$$

and whose discrimination parameters, $a_g$, and difficulty parameters, $b_g$, form nine different sets as shown in Table 1. Let us assume that ten subjects, whose ability levels are −1.52, −1.02, −0.87, −0.46, −0.21, 0.15, 0.53, 0.89, 1.24 and 1.63, have taken the test, with the criterion test information 25.00, i.e., the standard error of estimation .20. Also assume that the number of items whose item parameters are one of the nine sets shown in Table 1 is so large that there is no chance that we will run out of items to present to a single subject from each category. Item 5, whose difficulty is intermediate, was used as the initial item, and, using the monte carlo method, a response was given to the item. If the response was correct, then item 9 was given repeatedly until an incorrect answer was given and the likelihood function provided a local maximum. If the response to item 5 was incorrect, then item 1 was given in the same way, until the likelihood function provided a local maximum. From then on, an item whose information is the greatest at the current maximum likelihood estimate was selected and presented to the subject each time, until the amount of test information reached 25.00. Table 2 presents the number of items used for each subject and the resulting maximum likelihood estimates obtained in two repeated sessions, using the same criterion, for each subject. Figure 4 presents the eventual amount of information in the vicinity of the true ability level of each subject for Session 1 (solid curve) and for Session 2 (dashed curve), and the similar amounts of information assuming that the presentation of new items is terminated with the criterion test information 20.00 and 16.00 in Session 2 respectively (dotted curves). Although there are some discrepancies in the amount of information when the same criterion test information is used, these two sessions can be considered weakly parallel.

TABLE 1

Item Parameters of Ten Hypothetical Binary Items

| Item Group | Discrimination Parameter $a_g$ | Difficulty Parameter $b_g$ |
|---|---|---|
| 1 | 1.20 | -2.00 |
| 2 | 1.60 | -1.50 |
| 3 | 2.00 | -1.00 |
| 4 | 1.40 | -0.50 |
| 5 | 1.80 | 0.00 |
| 6 | 1.30 | 0.50 |
| 7 | 1.70 | 1.00 |
| 8 | 1.90 | 1.50 |
| 9 | 1.50 | 2.00 |

TABLE 2

Maximum Likelihood Estimate of Ability and the Number of Items Used in Each of Two Sessions of Tailored Testing Using the Same Criterion, for Each of the Ten Subjects

| Subject | Ability Level | Session 1 | | Session 2 | |
|---|---|---|---|---|---|
| | | Number of Items | MLE | Number of Items | MLE |
| 1 | -1.52 | 17 | -1.40 | 20 | -1.40 |
| 2 | -1.02 | 12 | -1.00 | 15 | -1.40 |
| 3 | -0.87 | 14 | -0.77 | 16 | -0.82 |
| 4 | -0.46 | 16 | -0.48 | 13 | -0.71 |
| 5 | -0.21 | 15 | 0.01 | 14 | -0.01 |
| 6 | 0.15 | 18 | 0.32 | 15 | 0.26 |
| 7 | 0.53 | 20 | 0.48 | 17 | 0.74 |
| 8 | 0.89 | 16 | 1.06 | 14 | 1.21 |
| 9 | 1.24 | 18 | 1.26 | 14 | 1.21 |
| 10 | 1.63 | 16 | 1.51 | 16 | 1.17 |

The sample error variances, which are unbiased estimates of the population error variance for any population to which these subjects belong and for which this specific procedure of tailored testing can be applied, were .010 and .022 respectively for the two sessions, whereas the theoretical population error variance was 1/20, or .040.

In this example, the same set of items, or item library, was used in two sessions. We can use two

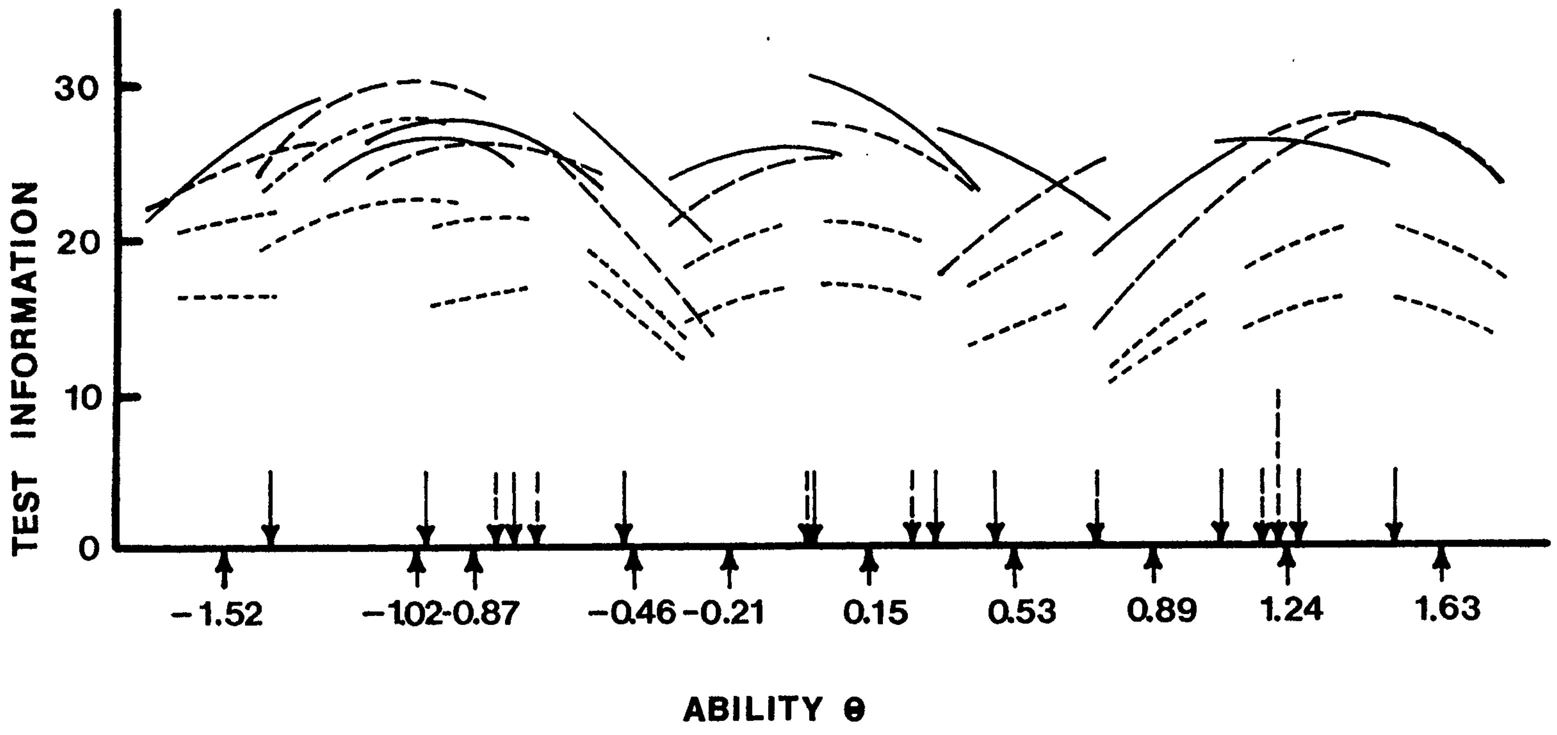


**Figure 4**

Amount of information at the termination of tailored testing in the vicinity of the true ability level of each of the ten subjects for Session 1 (solid curve) and for Session 2 (dashed curve), and similar curves assuming the presentation of new items was terminated when the amount of information at the current maximum likelihood estimate reached 20.00 and 16.00 respectively (dotted curves).

completely different sets of items in two sessions, however, for the purpose of repeating the measurement or giving two parallel tests, as long as these two sets of items are constructed to measure the same ability.

The way the local maximum likelihood estimate changes in both sessions after each presentation of a new item is shown for each of the ten subjects in the Appendix.

## Correlation Between the Repeated Testing Results

Since the standard error of estimation is defined independently of any specific group of subjects and can be considered solely as a property of the test or of the test procedure, as in tailored testing, this concept, instead of reliability, should be considered as fundamental in test theory. A strong merit of such a standard error of estimation is the fact that it is defined locally, or as a function of ability $\Theta$, unlike in classical test theory. For this reason, reliability is a dead concept in test theory since it differs from one group of subjects to another, and its generalizability is narrowly limited. Curiosity may prompt us however, to determine what the correlation coefficient would be if we administer tailored testing twice with the same procedure, or with two different procedures, to a specific group of people in whom we are particularly interested. If we can actually repeat the same procedure twice, or administer two different procedures, to the group of people, we can simply calculate the product-moment correlation coefficient between the two sets of $\hat{\Theta}$, and obtain the answer. The question is, therefore, can we estimate the correlation from the result of a single administration of a test? In fact, such an estimation is relatively easy, if the amount of information used as a criterion in terminating the testing is large enough to let us apply the normal approximation for the error distribution at every point of $\Theta$ under consideration. Under such conditions, by virtue of Equation 12, the covariance between ability $\Theta$ and error $\varepsilon$ is zero for any group of examinees (Samejima, 1976c), and the correlation coefficient between the two sets of maximum likelihood estimates is given by

$$\rho_{\hat{\theta}_1\hat{\theta}_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - E\{I(\theta)^{-1}\}]\ [\mathrm{Var.}\ (\hat{\theta}_1)]^{-1}\ . \qquad [16]$$

Thus the only term which we need to estimate is $E[I(\Theta)^{-1}]$. We could use, for instance, the mean of $I(\hat{\Theta}_1)^{-1}$ over all the examinees in a single administration of the test as an estimate of $E[I(\Theta)^{-1}]$.

It is interesting to note that such an estimate of the correlation coefficient is obtainable even if two different strategies or criteria are used in two tailored testing procedures. Suppose that we have actually administered the first tailored test with a certain strategy, and wish to know the correlation coefficient for the specific examinee group if we repeat the procedure with a different strategy. In such a non-parallel case, we obtain instead of Equation 16, the following:

$$\rho_{\hat{\theta}_1\hat{\theta}_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - E\{I^{(1)}(\theta)^{-1}\}]\ [\mathrm{Var.}\ (\hat{\theta}_1)\{\mathrm{Var.}\ (\hat{\theta}_1) - E\{I^{(1)}(\theta)^{-1}\} + E\{I^{(2)}(\theta)^{-1}\}\}]^{-1/2}, \qquad [17]$$

where $I^{(1)}(\Theta)$ is the criterion information function of the first testing, and $I^{(2)}(\Theta)$ is that of the second testing which is not actually conducted. Thus all we need to estimate are $E\{I^{(1)}(\Theta)^{-1}\}$ and $E\{I^{(2)}(\Theta)^{-1}\}$, and again we could use the means of $I^{(1)}(\hat{\Theta}_1)^{-1}$ and $I^{(2)}(\hat{\Theta}_1)^{-1}$ respectively.

If the criterion information function is constant for the entire range of $\Theta$ under consideration, then we have

$$E\{I^{(1)}(\theta)^{-1}\} = I^{(1)}(\theta)^{-1} = \sigma_1^2 \qquad [18]$$

and

$$E\{I^{(2)}(\theta)^{-1}\} = I^{(2)}(\theta)^{-1} = \sigma_2^2\ . \qquad [19]$$

Thus we can rewrite Equation 17 in the form

$$\rho_{\hat{\theta}_1\hat{\theta}_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - \sigma_1^2]$$

$$[\mathrm{Var.}\ (\ddot{\theta}_1)\{\mathrm{Var.}\ (\hat{\theta}_1) - \sigma_1^2 + \sigma_2^2\}]^{-1/2} \quad . \qquad [20]$$

In a special case where $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we have

$$\rho_{\hat{\theta}_1\hat{\theta}_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - \sigma^2][\mathrm{Var.}\ (\hat{\theta}_1)]^{-1}, \qquad [21]$$

the replacement for Equation 16. Clearly, all the necessary information in calculating the correlation coefficient through Equations 20 or 21 is observable in these simplified situations.

For the purpose of illustration, the product-moment correlation coefficient between the two sets of maximum likelihood estimates given in Table 2 was calculated, and it was .987. The sample variances of $\hat{\Theta}_1$ and $\hat{\Theta}_2$ were .891949 and 1.000265 respectively, and, using 0.04 for $\sigma^2$ in Equation 21, the estimated correlation coefficient is .955; the similar estimate obtained from the set of $\hat{\Theta}_2$ is .960. The reason the actual correlation is higher than these two estimated values is that the sample error variances, which were presented in the preceding section, are .010 and .022 for the two sessions; both are less than .040. In fact, if these values are used instead of .040, the estimated correlation coefficients, though not observable with actual data, are .989 and .978 respectively, which are closer to the actual correlation, .987.

If we use the mean of $I(\hat{\Theta})^{-1}$ as the estimate of $E[I(\Theta)^{-1}]$ in Equation 16, these values are .038006 and .038708 for the two sessions respectively, and the estimated correlation coefficients through Equation 16 are .957 and .961. Since our data are simulated data, we can compute the mean of $I(\Theta)^{-1}$ for each session, and with these values, .038680 and .041667, the estimated correlation coefficients through Equation 16 are .957 and .958 respectively, although they are not observable with actual data. These two sets of results are practically the same as the set of results obtained by using $\sigma^2 = 0.040$ in Equation 16, i.e., 0.955 and 0.960.

Figure 4 presents the amounts of test information assuming the criterion test information is 20.00 and 16.00 respectively, in the vicinity of the ability level of each subject. The maximum likelihood estimates for the ten examinees for the two situations in Session 2 are shown in Table 3. The actual correlation coefficient between each of these two sets of maximum likelihood estimates and the set of $\hat{\Theta}_1$ given in Table 2 was computed, and these values were .987 and .992 respectively. The estimated correlation coefficient using the set of $\hat{\Theta}_1$ given in Table 2 and .040 and .050 for the estimates of $\sigma_1^2$ and $\sigma_2^2$ respectively in Equation 20 were .950, and the similar estimate using the set of ten observations in the third column of Table 3 and reversing the estimates of $\sigma_1^2$ and $\sigma_2^2$ was .959. The corresponding estimated correlation coefficients for the case in which the criterion information in Session 2 is 16.00 instead of 20.00 were .943 and .956. Again these estimated values are less than the actual correlation coefficients, i.e., .950, .959 as opposed to .987, and .943, .956 against .992. The sample error variances were .072 for the case in which the criterion test information is 20.00, and 0.062 when it is 16.00. The means of

TABLE 3

Maximum Likelihood Estimate of Ability and the Number of Items Used in Session 2, When Each of the Two Criterion Informations, 20.00 and 16.00, Is Used

| Subject | Criterion: 20 | | Criterion: 16 | |
|---|---|---|---|---|
| | Number of Items | MLE | Number of Items | MLE |
| 1 | 18 | -1.49 | 15 | -1.49 |
| 2 | 14 | -1.37 | 12 | -1.48 |
| 3 | 14 | -0.64 | 12 | -0.72 |
| 4 | 11 | -0.63 | 10 | -0.68 |
| 5 | 12 | -0.01 | 10 | -0.01 |
| 6 | 12 | 0.27 | 10 | 0.16 |
| 7 | 15 | 0.83 | 12 | 0.68 |
| 8 | 11 | 1.35 | 10 | 1.25 |
| 9 | 11 | 1.35 | 9 | 1.50 |
| 10 | 13 | 1.30 | 11 | 1.43 |

$I(\hat{\Theta})^{-1}$ for the two cases were .046901 and .058462 respectively. Using each of them and .038680, the corresponding value for the result of Session 1, the two estimates of the correlation coefficient through Equation 17 were computed for each of the two cases, treating Session 1 as the first testing and then Session 2 as the first testing in each case. These estimates were .952 and .961 for the case in which 20.00 is used as the criterion test information, and .946 and .958 for the case where 16.00 is used. Again these values are very close to those obtained through Equation 20, i.e., .950 and .959 for the first case, and .943 and .956 for the second case. The means of $I(\Theta)^{-1}$ for the two cases, which are not observable with actual data, were .051190 and .063070. These values are still close enough to .0500 and .0625, which are used for the criteria for termination of testing in these two cases, i.e., $[20.00]^{-1}$ and $[16.00]^{-1}$, in spite of the fact that the values of criterion test information are less and, therefore, the distance from the true ability level to the maximum likelihood estimate, at which the amount of information was measured for the criterion purpose, is great with high probabilities.

## Discussion

It has been approximately thirty years since psychometricians initiated the idea of item-oriented test theory using the item characteristic function, and yet the theory has rarely been studied or appreciated by people in the areas of applied psychology and educational measurement, in spite of its superiority over classical test theory and its broader area of applicability. One of the main reasons for this unfortunate fact may be the mathematical complexity of the theory in comparison with classical test theory, and the only solution for this problem may be to try to invite people with a stronger mathematical background to these areas. On the other hand, since some applied areas like tailored testing require better foundations than classical test theory can provide, there has been a tendency among these researchers to try to adopt the theory. The information function is apparently one of the difficult topics in latent trait theory, judging from various questions and comments the author has personally received. However, it is one of the most useful concepts in the theory, and it is desirable that it be used in its full power. In the present paper, the attempt has been made to clarify some useful implications of the information function, and to suggest a way of using it in actual tailored testing.

## References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, (Eds.), *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Cramer, H. *Mathematical methods of statistics.* Princeton, New Jersey: Princeton University Press, 1946.

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

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17, 1969.

Samejima, F. A general model for free-response data. *Psychometrika Monograph,* No. 18, 1972.

Samejima, F. Homogeneous case of the continuous response model. *Psychometrika,* 1973, *38,* 203–219. (a)

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

Samejima, F. *Graded response model of the latent trait theory and tailored testing.* Paper presented at the Conference on Computerized Adaptive Testing, 1975, Civil Service Commission and Office of Naval Research, Washington, D.C.

Samejima, F. Effects of individual optimization in setting the boundaries of dichotomous items on accuracy of estimation. *Applied Psychological Measurement,* 1976, *1,* 77-94. (a)

Samejima F. *A way of estimating item characteristic functions.* Paper presented at the Joint Meeting of the Psychometric Society and the Mathematical Psychology Group, Bell Laboratories, Murray Hill, New Jersey, 1976. (b)

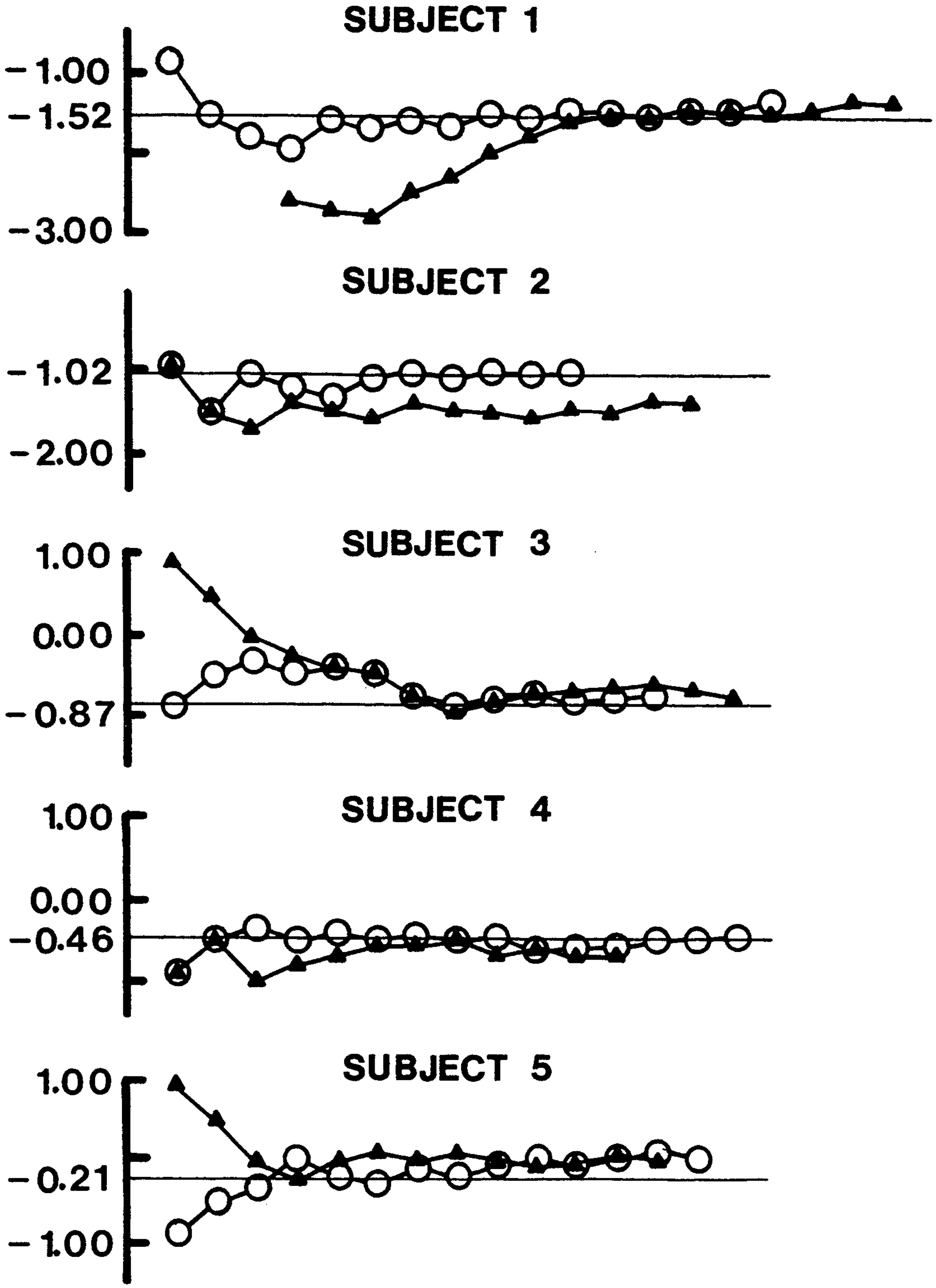
Samejima, F. *Weak parallel forms and a method of estimating item characteristic functions using the maximum likelihood estimate of ability.* In preparation, 1976. (c)

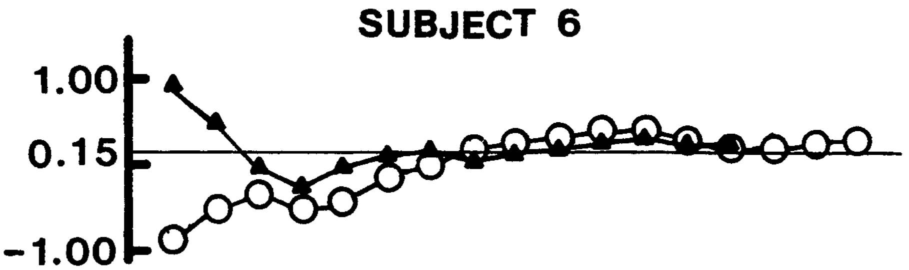
## Author's Address

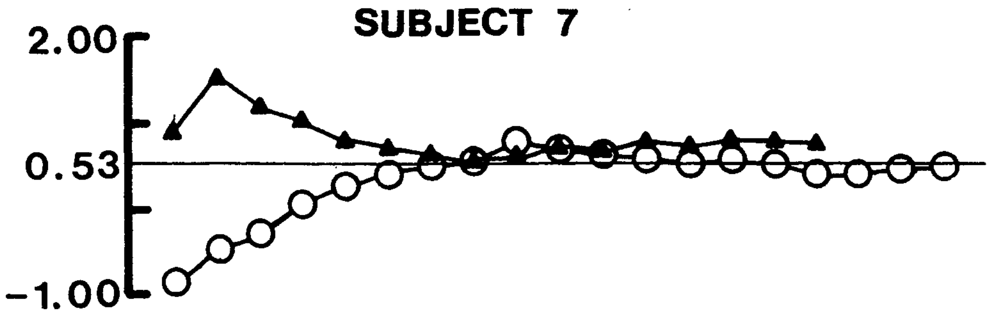
Fumiko Samejima, Department of Psychology, University of Tennessee, Knoxville TN 37916

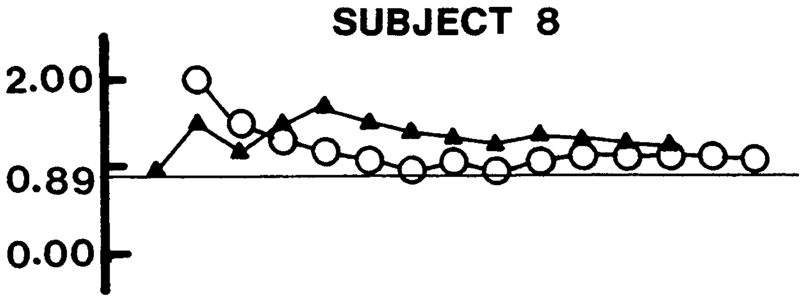
## Appendix

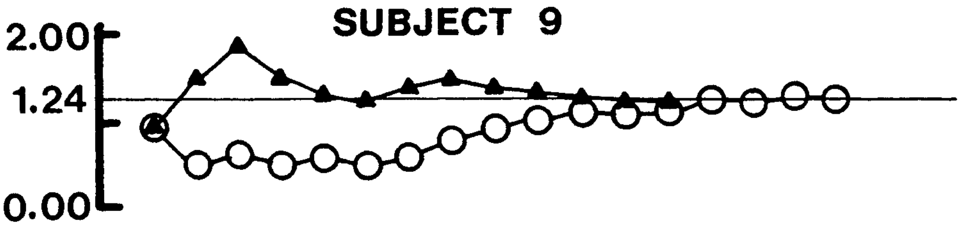
Graphic presentation of the change of the local maximum likelihood estimate after the presentation of each new item for each of the ten subjects.

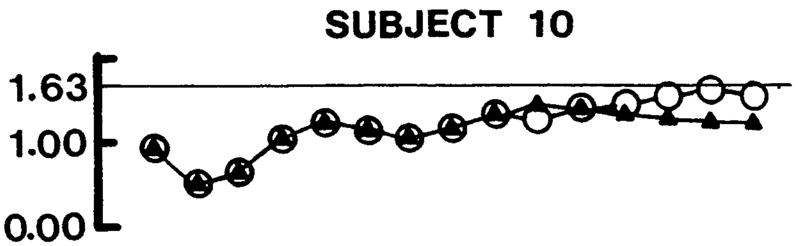



SUBJECT 6


SUBJECT 7


SUBJECT 8


SUBJECT 9


SUBJECT 10