Estimation of Reliability Coefficients Using the Test Information Function and Its Modifications

Fumiko Samejima
University of Tennessee

The reliability coefficient and the standard error of measurement in classical test theory are not properties of a specific test, but are attributed to both a specific test and a specific trait distribution. In latent trait models, or item response theory, the test information function (TIF) provides more precise local measures of accuracy in trait estimation than are available from the reliability coefficient. The reliability coefficient is still widely used, however, and is popular because of its simplicity. Thus, it is worthwhile to relate it to the TIF. In this paper, the reliability coefficient is predicted from the TIF, or two modified TIF formulas, and a specific trait distribution. Examples demonstrate the variability of the reliability coefficient across different trait distributions, and the results are compared with empirical reliability coefficients. Practical suggestions are given as to how to make better use of the reliability coefficient. *Index terms: adaptive testing, bias, classical test theory, item information function, latent trait models, maximum likelihood estimation, reliability coefficient, standard error of measurement, test information function, trait estimation.*

Reliability and validity coefficients have been widely accepted by psychologists and test users as important concepts in classical test theory (CTT). The reliability coefficient ($r_{X_1X_2}$), where X_1 is the test score and X_2 is the retest score, can be expressed as the ratio of the true score variance to the observed test score variance. $r_{X_1X_2}$ is largely influenced by the homogeneity or heterogeneity of the group of examinees to whom the test was administered as well as properties of the test. In spite of this, many researchers and test users still treat $r_{X_1X_2}$ as if it were an attribute solely of the test.

In latent trait models, or item response theory (IRT), the item information function (IIF) and the test information function (TIF) provide measures of the local accuracy of trait estimation, a concept that is missing in CTT. The values of the IIF and the TIF do not depend on the specific group of examinees tested, unlike $r_{X_1X_2}$ (i.e., the IIF and TIF are population-free). Therefore, $r_{X_1X_2}$ and the associated standard error of measurement (SEM) are not as important as they were before IRT became feasible. However, because of its simplicity $r_{X_1X_2}$ is still popular among test users. Thus, it is worthwhile to relate it to the TIF, and to suggest better ways to use $r_{X_1X_2}$.

Lawley (1943) related the normal ogive model and $r_{X_1X_2}$. He showed that $r_{X_1X_2}$ is obtained from the estimated error score variance and the observed test score variance when the trait distribution is N(0,1) and the difficulty parameters of the n dichotomous test items distribute normally. Lord (1952) showed how the discrimination index of a test in CTT at a given trait level (which is closely related to the amount of test information) is related to $r_{X_1X_2}$ at a specified trait level, when the regression of test score on θ level is approximately linear. Samejima (1977b) showed that $r_{X_1X_2}$ can be obtained from the TIF and the trait distribution of a target population. Lord (1983) discussed unbiased estimators of the parallel forms $r_{X_1X_2}$ in the three-parameter logistic model. The idea of replacing the CTT concepts of $r_{X_1X_2}$ and the SEM by the TIF in IRT is advanced by the proposal of two modified TIF formulas (Samejima, 1990), which use the bias function of the maximum likelihood estimate (MLE) (Samejima, 1987, 1993a, 1993b).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 3, September 1994, pp. 229–244

0146-6216/94/030229-16$2.05

The present paper uses information from the MLE bias function (MLEBF) of a test to predict the $r_{X_1X_2}$ and SEM attributed to a specified examinee group to whom a test is to be administered. The results of predicting $r_{X_1X_2}$ using the original TIF and the two modified versions of the TIF also are compared.

The TIF and Local Standard Error of Estimation

Let θ be the latent trait that takes on any real number. Assume that there is a set of n test items measuring θ whose characteristics are known. Let g denote such an item, k_g be a discrete response to item g, and $P_{k_g}(\theta)$ denote the operating characteristic of k_g, or the conditional probability assigned to k_g, given θ; that is,

$$P_{k_g}(\theta) = \text{Prob}\left[k_g|\theta\right]. \tag{1}$$

Assume that $P_{k_g}(\theta)$ is at least five-times differentiable with respect to θ. The item response information function (Samejima, 1972) is defined as

$$I_{k_g}(\theta) = -\frac{\partial^2}{\partial\theta^2}\log P_{k_g}(\theta), \tag{2}$$

and the IIF, $I_g(\theta)$, is defined as the conditional expectation of $I_{k_g}(\theta)$, given θ, such that

$$I_g(\theta) = \text{E}\left[I_{k_g}(\theta)|\theta\right] = \sum_{k_g} I_{k_g}(\theta)P_{k_g}(\theta). \tag{3}$$

When item g is scored dichotomously, the IIF is simplified to

$$I_g(\theta) = \left[\frac{\partial}{\partial\theta}P_g(\theta)\right]^2\left\{\left[P_g(\theta)\right]\left[1-P_g(\theta)\right]\right\}^{-1}, \tag{4}$$

where $P_g(\theta)$ is the operating characteristic of the correct answer to item g. This is identical to the item information function proposed by Birnbaum (1968) for the dichotomous response item. Let $\mathbb{V}$ be a response pattern such that

$$\mathbb{V} = \left\{k_g\right\}' \quad g = 1, 2, \ldots, n. \tag{5}$$

The operating characteristic, $P_{\mathbb{V}}(\theta)$, of the response pattern $\mathbb{V}$ is defined as the conditional probability of $\mathbb{V}$, given θ, and assuming local independence (Lord & Novick, 1968),

$$P_{\mathbb{V}}(\theta) = \prod_{k_g\in\mathbb{V}} P_{k_g}(\theta). \tag{6}$$

The response pattern information function (Samejima, 1972), $I_{\mathbb{V}}(\theta)$, is given by

$$I_{\mathbb{V}}(\theta) = -\frac{\partial^2}{\partial\theta^2}\log P_{\mathbb{V}}(\theta) = \sum_{k_g\in\mathbb{V}} I_{k_g}(\theta). \tag{7}$$

The TIF, $I(\theta)$, is defined as the conditional expectation of $I_{\mathbb{V}}(\theta)$, given θ, and from Equations 2, 3, 6, and 7,

$$I(\theta) = \text{E}\left[I_{\mathbb{V}}(\theta)|\theta\right] = \sum_{\mathbb{V}} I_{\mathbb{V}}(\theta)\,P_{\mathbb{V}}(\theta) = \sum_{g=1}^{n} I_g(\theta). \tag{8}$$

Again, the relationship between the IIF and the TIF demonstrated in Equation 8 is identical to the result Birnbaum (1968) demonstrated at the dichotomous response level.

The reciprocal of the square root of the TIF, $[I(\theta)]^{-1/2}$, is the asymptotic standard deviation of the conditional distribution of the MLE of θ, given its true value. This function usually is used as the standard error of estimation (SEE) even when the number of test items is finite and relatively small. Note that the SEE is a function of θ--it is locally defined. Also, unlike its counterpart in CTT, the SEE does not depend on a

specific group of examinees, but is solely a property of the test.

In CTT, the SEE gives the impression that the test always provides the extent of error indicated by its value. Common sense suggests, however, that no test can be appropriate for every group of examinees. For example, if a simple arithmetic addition test is administered to college mathematics majors, almost everyone will correctly answer all the items. In such a case, the test is useless, because the results do not reflect the individual differences among the examinees. If a calculus test is administered to elementary school students, opposite—but similar—results will occur. Thus each test is effective only locally on the θ dimension. The SEM of the test should differ, therefore, for groups of examinees at different locations on the θ scale. It is more appropriate to consider the error of measurement as a function of θ, as IRT models do.

Prediction of the Reliability Coefficient and the SEM for a Specific θ Distribution Using the TIF

Using the TIF, it is possible to link CTT with IRT through the prediction of $r_{X_1X_2}$ and the SEM for a specified θ distribution or a specified group of examinees (Samejima, 1977b).

Let θ_V^* be any estimator of θ,

$$\theta_V^* = \theta + \varepsilon, \tag{9}$$

where ε denotes the error variable. In the test-retest situation,

$$\begin{cases} \theta_{V1}^* = \theta + \varepsilon_1 \\ \theta_{V2}^* = \theta + \varepsilon_2, \end{cases} \tag{10}$$

where a subscript 1 indicates the test, and a subscript 2 indicates the retest. If it can be assumed that in the test-retest situation,

$$\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0, \tag{11}$$

$$\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2), \tag{12}$$

and

$$\mathrm{Cov}(\theta, \varepsilon_1) = \mathrm{Cov}(\theta, \varepsilon_2) = 0, \tag{13}$$

then

$$\mathrm{Cov}(\theta_{V1}^*, \theta_{V2}^*) = \mathrm{Var}(\theta) = \mathrm{Var}(\theta_V^*) - \mathrm{Var}(\varepsilon), \tag{14}$$

where θ_V^* represents either θ_{V1}^* or θ_{V2}^*, and ε represents either ε_1 or ε_2. Thus, the correlation between θ_{V1}^* and θ_{V2}^* is

$$r(\theta_{V1}^*, \theta_{V2}^*) = \left[\mathrm{Var}(\theta_V^*) - \mathrm{Var}(\varepsilon)\right]\left[\mathrm{Var}(\theta_V^*)\right]^{-1}. \tag{15}$$

Note that if θ is replaced by one of its transformed forms (i.e., the true test score T), and if the observed test score X is used as the estimator of T, and E is used as its error of estimation, then Equation 9 can be rewritten as

$$T = X + E, \tag{16}$$

then Equation 15 becomes a familiar equation for the reliability coefficient $r_{X_1X_2}$,

$$r_{X_1X_2} = \mathrm{Var}(T)\left[\mathrm{Var}(X)\right]^{-1}. \tag{17}$$

In general,

$$\mathrm{Var}(\varepsilon) = \mathrm{E}\left[\varepsilon - \mathrm{E}(\varepsilon)\right]^2 = \mathrm{E}\left[\varepsilon - \mathrm{E}(\varepsilon|\theta)\right]^2 + \mathrm{E}\left[\mathrm{E}(\varepsilon|\theta) - \mathrm{E}(\varepsilon)\right]^2 + 2\mathrm{E}\left\{\left[\varepsilon - \mathrm{E}(\varepsilon|\theta)\right]\left[\mathrm{E}(\varepsilon|\theta) - \mathrm{E}(\varepsilon)\right]\right\}. \tag{18}$$

If the error variable ε is conditionally unbiased for the θ interval of interest, then Equation 18 will be

reduced to

$$\mathrm{Var}(\varepsilon) = \mathrm{E}\left[\varepsilon^2\right]. \tag{19}$$

When the MLE of θ is used, let $\hat{\theta}_V$ or $\hat{\theta}$ denote the MLE of θ based on the response pattern V. Samejima (1977a, 1979) observed that even with a relatively small number of test items the conditional distribution of $\hat{\theta}$, given θ, can be approximated by the normal distribution $N\{\theta,[I(\theta)]^{-1/2}\}$ if two conditions hold. Condition 1 is that $\hat{\theta}$ must be practically conditionally unbiased for the θ interval of interest. Condition 2 is that $I(\theta)$ must assume reasonably high values for that specific interval. The best approximation occurs when this interval covers the range of θ within which most examinees are located. When this is the case, from Equation 19,

$$\mathrm{Var}(\varepsilon) \doteq \mathrm{E}\left\{\left[I(\theta)\right]^{-1}\right\} = \int_{-\infty}^{\infty}\left[I(\theta)\right]^{-1} f(\theta)d\theta, \tag{20}$$

where $f(\theta)$ denotes the density function of θ for a specific group of examinees. Thus, from Equation 15,

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left\{\mathrm{Var}\left(\hat{\theta}\right) - \mathrm{E}\left\{\left[I(\theta)\right]^{-1}\right\}\right\}\left[\mathrm{Var}\left(\hat{\theta}\right)\right]^{-1}. \tag{21}$$

Equation 21 indicates that the reliability coefficient $\left[r\left(\hat{\theta}_1, \hat{\theta}_2\right)\right]$ can be predicted by a single administration of the test, given $I(\theta)$ and the θ distribution of the examinees.

It also has been observed (Samejima, 1977b) that in computerized adaptive testing (CAT), $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ can be predicted if a specified amount of test information is used as the stopping rule for a given θ level in the test and retest situations. Thus,

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left\{\mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{E}\left\{\left[I_{(1)}(\theta)\right]^{-1}\right\}\right\}\left\{\mathrm{Var}\left(\hat{\theta}_1\right)\left\{\mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{E}\left\{\left[I_{(1)}(\theta)\right]^{-1}\right\} + \mathrm{E}\left\{\left[I_{(2)}(\theta)\right]^{-1}\right\}\right\}\right\}^{-1/2}, \tag{22}$$

where $I_{(1)}(\theta)$ and $I_{(2)}(\theta)$ are the preset criterion TIFs in the test and retest situations, respectively, which are adopted as the stopping rules for the two testing sessions (a subscript 1 indicates the test situation, and a subscript 2 indicates the retest situation). Note that these two criterion TIFs need not be the same for the test and the retest, nor do they need to be constant for all θ. Note also that $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ is obtainable from a single test administration, because all that is needed is the sample variance of $\hat{\theta}_i$ to replace $\mathrm{Var}(\hat{\theta}_1)$ in Equation 22. $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ will change if the preset criterion TIFs [$I_{(1)}(\theta)$, $I_{(2)}(\theta)$, or both] change. For the simplified case in which the same amount of test information is used as the criterion for terminating the presentation of new items for every examinee in each of the test and retest situations, respectively, Equation 22 can be rewritten as

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left[\mathrm{Var}\left(\hat{\theta}_1\right) - \sigma_1^2\right]\left\{\mathrm{Var}\left(\hat{\theta}_1\right)\left[\mathrm{Var}\left(\hat{\theta}_1\right) - \sigma_1^2 + \sigma_2^2\right]\right\}^{-1/2}, \tag{23}$$

where σ_1^2 and σ_2^2 are the reciprocals of the constant amounts of criterion test information in the two testing situations, respectively. If a constant amount of test information is used as the stopping rule for every examinee in both the test and retest situations, then $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ takes the simplest form

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left[\mathrm{Var}\left(\hat{\theta}_1\right) - \sigma^2\right]\left[\mathrm{Var}\left(\hat{\theta}_1\right)\right]^{-1}, \tag{24}$$

where σ^2 denotes the reciprocal of this common constant amount of test information.

The appropriateness of the above normal approximation of the conditional distribution of $\hat{\theta}$, given θ, can be examined using the monte carlo method (see Samejima, 1977a). A necessary condition for this approximation is that $\hat{\theta}$ is practically conditionally unbiased, that is, the regression of $\hat{\theta}$ on θ is very close to θ itself, for the interval of interest. This can be examined using the MLEBF of the test, which is intro-

duced below. Note that the MLEBF together with the θ distribution of the target population also determines whether the assumption described by Equation 13 can be accepted.

Because the SEM in CTT is for the entire group of examinees, the SEM, σ_ε, of a test attributed to a specific θ distribution can be written as

$$\sigma_\varepsilon = \mathrm{E}\left\{\left[I(\theta)\right]^{-1/2}\right\} = \int_{-\infty}^{\infty}\left[I(\theta)\right]^{-1/2} f(\theta)d\theta\,, \tag{25}$$

when Conditions 1 and 2 described above hold.

Assume that the test items have already been calibrated; that is, the item parameters of each item have been estimated following an appropriate mathematical model for $P_{k_g}(\theta)$. This can be done either separately from, or simultaneously with, maximum likelihood estimation of the individual parameter θ for a nonadaptive test. In the former situation, $\hat{\theta}_V$ will be obtained for every individual by using the estimated item parameters. In CAT, the item parameters for the items in the item pool usually have been estimated before using the item pool for adaptive testing; using these estimated item parameters the individual parameter $\hat{\theta}_V$ is obtained as a result of testing, which is based on a tailored subset of the item pool for each examinee. Thus, $\mathrm{Var}(\hat{\theta})$ and $\mathrm{Var}(\hat{\theta}_1)$ in Equations 21 or 22 can be computed.

From Equations 2, 3, and 8, $I(\theta)$ for a nonadaptive test can be estimated using the estimated item parameters. The density function, $f(\theta)$, can either be estimated (e.g., Samejima, 1981) or set a priori if prediction is the purpose, as it is here. Thus, $\mathrm{E}\{[I(\theta)]^{-1}\}$ is obtained with $I(\theta)$ replaced by the estimated TIF in Equation 20, approximating the area by a number of rectangles of small widths or following Simpson's quadrature formula (Elderton & Johnson, 1969) for the approximation to the integration. Using the result of Equation 20 and the computed sample variance of $\hat{\theta}_1$ that replaces $\mathrm{Var}(\hat{\theta})$ in Equation 21, the $r(\hat{\theta}_1, \hat{\theta}_2)$ attributed both to the test and to the specific examinee group is obtained. Similarly, the SEM can be obtained by Equation 25. For a CAT, $r(\hat{\theta}_1, \hat{\theta}_2)$ can be obtained by using the preset criteria—$I_{(1)}(\theta)$ for the test and $I_{(2)}(\theta)$ for the retest—and the sample variance of $\hat{\theta}_1$ that replaced $\mathrm{Var}(\hat{\theta}_1)$ in Equation 22.

If conditional unbiasedness is not supported for the θ interval of interest, appropriate modifications for Equations 21 and 25 are needed. This can be done using the MLEBF (Lord, 1983; Samejima, 1987, 1993a, 1993b) and the modified TIFs (Samejima, 1990).

The MLEBF and Two Modified TIF Formulas

Lord (1983) proposed a bias function for the MLE of θ in the three-parameter logistic model; its operating characteristic for the correct response, $P_g(\theta)$, is given by

$$P_g(\theta) = c_g + (1 - c_g)\left\{1 + \exp\left[-\mathrm{D}a_g(\theta - b_g)\right]\right\}^{-1}, \tag{26}$$

where a_g, b_g, and c_g are the item discrimination, item difficulty, and guessing parameters, respectively, and D is a scaling factor, which is set equal to 1.7 when the logistic model is used as a substitute for the normal ogive model. Lord's bias function, denoted by $B(\theta; \hat{\theta}_V)$, can be written as

$$B(\theta; \hat{\theta}_V) = \mathrm{D}\left[I(\theta)\right]^{-2}\sum_{g=1}^{n} a_g I_g(\theta)\left[\Psi_g(\theta) - \frac{1}{2}\right], \tag{27}$$

where

$$\Psi_g(\theta) = \left\{1 + \exp\left[-\mathrm{D}a_g(\theta - b_g)\right]\right\}^{-1}. \tag{28}$$

In Equation 27, the bias should be negative when $\Psi_g(\theta)$ is less than .5 for all the items [which is necessarily the case for some interval of θ, $(-\infty, \theta_L)$], and should be positive when $\Psi_g(\theta)$ is greater than .5 for all

items [which also necessarily happens for some interval, $(\theta_H, +\infty)$], and in between these values of θ the bias tends to be close to 0.0. This is obvious because the last factor on the right-hand side of Equation 27 assumes negative values for some items and positive values for others, and their total tends to be close to 0.0, provided that the b_gs are distributed over a wide range of θ. Lord (1984) applied this MLEBF to an 85-item verbal test and found that the bias was practically 0.0 for a wide range of θ.

Samejima (1987, 1993a, 1993b) expanded the MLEBF to include any discrete item responses. The MLEBF in the general case can be written as

$$B\left(\theta;\hat{\theta}_V\right) = E\left(\hat{\theta}_V - \theta|\theta\right) = -(1/2)\left[I(\theta)\right]^{-2}\sum_{g=1}^{n}\sum_{k_g} A_{k_g}(\theta)P''_{k_g}(\theta) = -(1/2)\left[I(\theta)\right]^{-2}\sum_{g=1}^{n}\sum_{k_g} P'_{k_g}(\theta)P''_{k_g}(\theta)\left[P_{k_g}(\theta)\right]^{-1}, \quad (29)$$

where $A_{k_g}(\theta)$ is the basic function (Samejima, 1969) for the discrete item response k_g, and $P'_{k_g}(\theta)$ and $P''_{k_g}(\theta)$ denote the first and second partial derivatives of $P_{k_g}(\theta)$ with respect to θ, respectively. For the graded response model in which item score x_g assumes successive integers, 0 through m_g, each k_g in Equation 29 must be replaced by the graded item score x_g. For a dichotomous response model, it can be reduced to the form

$$B\left(\theta;\hat{\theta}_V\right) = E\left(\hat{\theta}_V - \theta|\theta\right) = -(1/2)\left[I(\theta)\right]^{-2}\sum_{g=1}^{n} I_g(\theta)P''_g(\theta)\left[P'_g(\theta)\right]^{-1}, \quad (30)$$

where $P'_g(\theta)$ and $P''_g(\theta)$ indicate the first and second partial derivatives of $P_g(\theta)$ with respect to θ, respectively. Equation 30 includes Lord's bias function in the three-parameter logistic model as a special case.

Samejima (1990) proposed two modified formulas for the TIF. They both use the MLEBF. One modified version takes the reciprocal of an approximate minimum variance bound, and the other modified version takes an approximate minimum bound of the mean squared error of the maximum likelihood estimator. The first modified TIF, $\Upsilon(\theta)$, is defined by

$$\Upsilon(\theta) = I(\theta)\left[1 + \frac{\partial}{\partial\theta}B\left(\theta;\hat{\theta}_V\right)\right]^{-2}. \quad (31)$$

The first partial derivative of the MLEBF with respect to θ, which is used in Equation 31, is provided by

$$\frac{\partial}{\partial\theta}B\left(\theta;\hat{\theta}_V\right) = \left[I(\theta)\right]^{-1}\left\{(1/2)\left[I(\theta)\right]^{-1}\sum_{g=1}^{n}\sum_{k_g}\left\{I_{k_g}(\theta)P''_{k_g}(\theta) - P'_{k_g}(\theta)P'''_{k_g}(\theta)\left[P_{k_g}(\theta)\right]^{-1}\right\} - 2B\left(\theta;\hat{\theta}_V\right)I'(\theta)\right\} \quad (32)$$

for the general case of discrete item responses, where $P'''_{k_g}(\theta)$ and $I'(\theta)$ denote the third and the first partial derivatives of $P_{k_g}(\theta)$ and $I(\theta)$ with respect to θ, respectively. For a set of dichotomous items, Equation 32 becomes

$$\frac{\partial}{\partial\theta}B\left(\theta;\hat{\theta}_V\right) = \left[I(\theta)\right]^{-1}\left\{(1/2)\left[I(\theta)\right]^{-1}\sum_{g=1}^{n}\left[P_g(\theta)\right]^{-2}\left[1-P_g(\theta)\right]^{-2}\right.$$
$$\left.\left\{\left[1-2P_g(\theta)\right]\left[P'_g(\theta)\right]^2P''_g(\theta) - P_g(\theta)\left[1-P_g(\theta)\right]\left\{\left[P''_g(\theta)\right]^2 + P'_g(\theta)P'''_g(\theta)\right\}\right\} - 2B\left(\theta;\hat{\theta}_V\right)I'(\theta)\right\}, \quad (33)$$

where $P'''_g(\theta)$ indicates the third partial derivative of $P_g(\theta)$ with respect to θ (see Samejima, 1987, 1990, 1993a, 1993b).

Equation 31 shows that the relationship between this new function and the original TIF depends on the first partial derivative of the MLEBF. To be more precise, if the partial derivative is positive, $\Upsilon(\theta)$ will be less than $I(\theta)$; if it is negative, this relationship will be reversed; if it is 0.0 (i.e., if the MLE is conditionally unbiased), $\Upsilon(\theta)$ and $I(\theta)$ will have the same value.

The second modified TIF, $\Xi(\theta)$, is defined by

$$\Xi(\theta) = I(\theta)\left\{\left[1+\frac{\partial}{\partial\theta}B\left(\theta;\hat{\theta}_V\right)\right]^2 + I(\theta)\left[B\left(\theta;\hat{\theta}_V\right)\right]^2\right\}^{-1}. \tag{34}$$

The difference between $\Upsilon(\theta)$ and $\Xi(\theta)$ (Equations 31 and 34, respectively), is the second and last term in the braces of the right-hand side of Equation 34. Because this term is non-negative throughout the entire range of θ, $\Xi(\theta) \leq \Upsilon(\theta)$ regardless of the slope of the MLEBF.

When the MLEBF of a test is monotonically increasing, Equations 31 and 34 show that $\Upsilon(\theta)$ and $\Xi(\theta)$ will never be larger than $I(\theta)$. In this specific case, $\Xi(\theta) \leq \Upsilon(\theta) \leq I(\theta)$ throughout the entire range of θ.

$r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ and the SEM When Conditional Unbiasedness of the MLE of θ Does Not Hold

When practical conditional unbiasedness of the MLE of θ does not exist—that is, $B\left(\theta;\hat{\theta}_V\right)$ is not approximately 0.0 for all values of θ in the interval of interest—$\Upsilon(\theta)$ or $\Xi(\theta)$ should be substituted for $I(\theta)$ in Equations 21 and 25. Thus, Equation 21 can be rewritten as

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left\{\mathrm{Var}\left(\hat{\theta}\right) - \mathrm{E}\left\{\left[\Upsilon(\theta)\right]^{-1}\right\}\right\}\left[\mathrm{Var}\left(\hat{\theta}\right)\right]^{-1} \tag{35}$$

or

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left\{\mathrm{Var}\left(\hat{\theta}\right) - \mathrm{E}\left\{\left[\Xi(\theta)\right]^{-1}\right\}\right\}\left[\mathrm{Var}\left(\hat{\theta}\right)\right]^{-1}. \tag{36}$$

In theory, $\Xi(\theta)$ is more appropriate, but in many cases discrepancies between $\Upsilon(\theta)$ and $\Xi(\theta)$ are small (see Samejima, 1990), so $\Upsilon(\theta)$ can be a good substitute for $\Xi(\theta)$. Also, in CAT, $\Upsilon(\theta)$ or $\Xi(\theta)$ can be used as the stopping rule in place of $I(\theta)$, and Equation 22 can be revised into the forms

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left\{\mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{E}\left\{\left[\Upsilon_{(1)}(\theta)\right]^{-1}\right\}\right\}\left\{\mathrm{Var}\left(\hat{\theta}_1\right)\left\{\mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{E}\left\{\left[\Upsilon_{(1)}(\theta)\right]^{-1}\right\} + \mathrm{E}\left\{\left[\Upsilon_{(2)}(\theta)\right]^{-1}\right\}\right\}\right\}^{-1/2} \tag{37}$$

or

$$r\left(\hat{\theta}_1, \hat{\theta}_2\right) = \left\{\mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{E}\left\{\left[\Xi_{(1)}(\theta)\right]^{-1}\right\}\right\}\left\{\mathrm{Var}\left(\hat{\theta}_1\right)\left\{\mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{E}\left\{\left[\Xi_{(1)}(\theta)\right]^{-1}\right\} + \mathrm{E}\left\{\left[\Xi_{(2)}(\theta)\right]^{-1}\right\}\right\}\right\}^{-1/2}. \tag{38}$$

In the same way, the two modified SEMs, $\sigma_{\varepsilon,1}$ and $\sigma_{\varepsilon,2}$, which are attributed to a specific distribution of θ, can be rewritten as

$$\sigma_{\varepsilon,1} = \mathrm{E}\left\{\left[\Upsilon(\theta)\right]^{-1/2}\right\} = \int_{-\infty}^{\infty}\left[\Upsilon(\theta)\right]^{-1/2} f(\theta)d\theta \tag{39}$$

and

$$\sigma_{\varepsilon,2} = \mathrm{E}\left\{\left[\Xi(\theta)\right]^{-1/2}\right\} = \int_{-\infty}^{\infty}\left[\Xi(\theta)\right]^{-1/2} f(\theta)d\theta. \tag{40}$$

Equation 29 shows that $B\left(\theta;\hat{\theta}_V\right)$ can be estimated if each item has been calibrated and the estimated $P_{k_g}(\theta)$ has been obtained. Thus, using this estimated MLEBF, $\Upsilon(\theta)$ and $\Xi(\theta)$ are estimated using Equations 31 and 34, respectively. Using these two modified TIFs, both $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ and the SEM can be obtained using Equations 35, 36, 39, and 40, in a similar manner as was described when $I(\theta)$ was used.

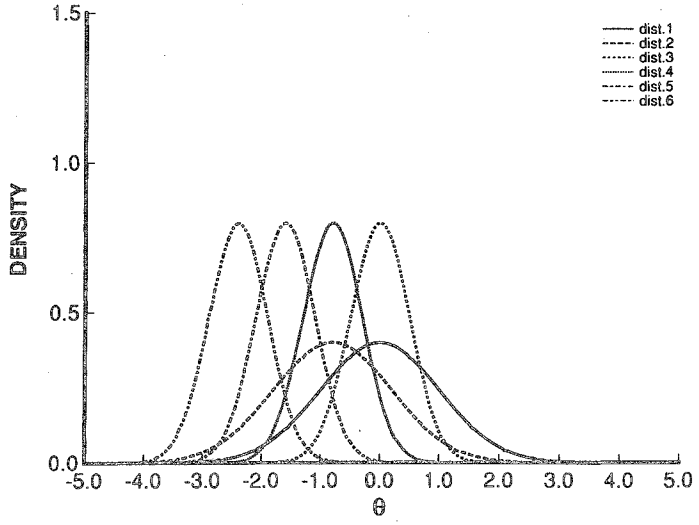
Example Applications

TIFs and Predicted $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$s and SEMs

Method. Six distributions of θ were hypothesized. Predictions were made for each of the six θ distributions about the $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$s and SEMs that would be obtained for a hypothetical test using the three different

TIF formulas. From Equations 21 and 25, the $r(\hat{\theta}_1, \hat{\theta}_2)$ and SEM were obtained using $I(\theta)$; from Equations 35 and 39, the $r(\hat{\theta}_1, \hat{\theta}_2)$ and SEM were obtained using $\Upsilon(\theta)$; and from Equations 36 and 40, the $r(\hat{\theta}_1, \hat{\theta}_2)$ and SEM were obtained using $\Xi(\theta)$. The six hypothetical distributions of θ were normally distributed with different means and standard deviations: Distribution 1, N(0, 1); Distribution 2, N(−.8, 1); Distribution 3, N(0, .5); Distribution 4, N(−.8, .5); Distribution 5, N(−1.6, .5); and Distribution 6, N(−2.4, .5), respectively. Figure 1 shows the density functions of these six distributions of θ.

Figure 1
Density Function of Six Hypothetical Distributions of θ:
N(0, 1), N(−.8, 1), N(0, .5), N(−.8, .5), N(−1.6, .5) and N(−2.4, .5)
[From Samejima (1994). Reprinted by Permission of Kluwer Academic Publishers]



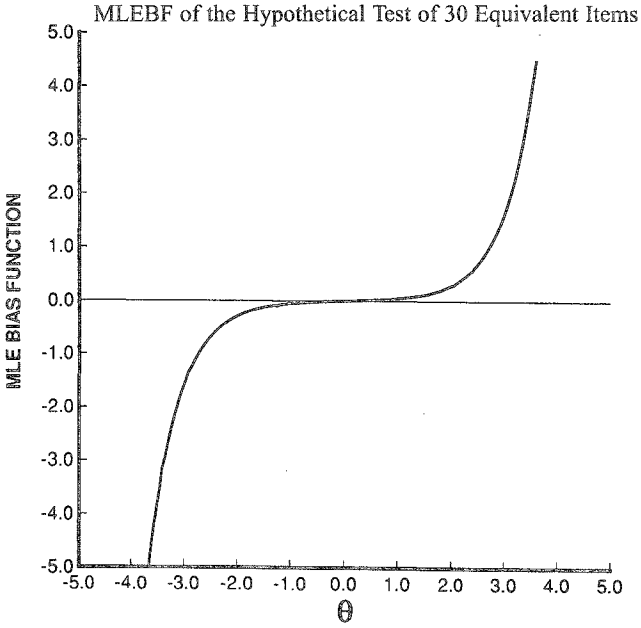
The hypothetical test consisted of 30 equivalent dichotomous items, which followed the logistic model (Equation 26) with $a_g = 1.0$, $b_g = 0.0$, $c_g = 0.0$, and D = 1.7. This particular hypothetical test was selected because the interval of θ for which practical conditional unbiasedness of the MLE $\hat{\theta}_V$, given θ, holds was expected to be small because of the common difficulty parameter for all the items; therefore, the discrepancies between $\Upsilon(\theta)$ or $\Xi(\theta)$ and $I(\theta)$ were expected to be large for a wider range of θ than those for a more typical test. This choice was made for the purpose of comparing the predictions of $r(\hat{\theta}_1, \hat{\theta}_2)$ for a specific distribution of θ when $I(\theta)$ was used versus $\Upsilon(\theta)$ or $\Xi(\theta)$.

Another reason for this choice was considerations in CAT. In CAT, except for the initial few items presented to an examinee, the tailored subset of items selected from the item pool consists of nearly equivalent items. If the same level of accuracy in estimating θ for all examinees is desired, for example, then it is reasonable to use a single specified amount of test information as the criterion in the stopping rule. In so doing, the choice between $I(\theta)$ and $\Upsilon(\theta)$ or $\Xi(\theta)$ will make a substantial difference, especially for examinees with very high levels of θ and for examinees with very low levels of θ, because in many cases the item pool will lack extremely difficult and extremely easy items.

Results. The MLEBF of this hypothetical 30-item test is shown in Figure 2. Note that outside the interval of θ (−1.0, 1.0) the amount of bias becomes increasingly large. The square roots of the TIFs [$I(\theta)$, $\Upsilon(\theta)$, and $\Xi(\theta)$] are shown in Figure 3.

Tables 1 and 2 present the predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s and SEMs for the six different distributions of θ, respectively. Integration in Equations 21, 35, and 36 was approximated by dividing θ into small steps of an

Figure 2
MLEBF of the Hypothetical Test of 30 Equivalent Items



equal width of .05 and using a number of rectangles. In Table 1, the mean and standard deviation (SD) of θ for each of the six distributions also are given. These SDs are slightly different from the squares of the second parameters of the normal distributions—.99157 versus 1.00000 for Distributions 1 and 2, and .50155 versus .50000 for Distributions 3, 4, 5, and 6, respectively, whereas all of the means are the same as the hypothesized normal distributions. The discrepancies in SDs are a result of using the rectangle method, which uses the density of each of the equally spaced points of θ as one side and the step width, .05, as the other, in order to approximate the normal distributions.

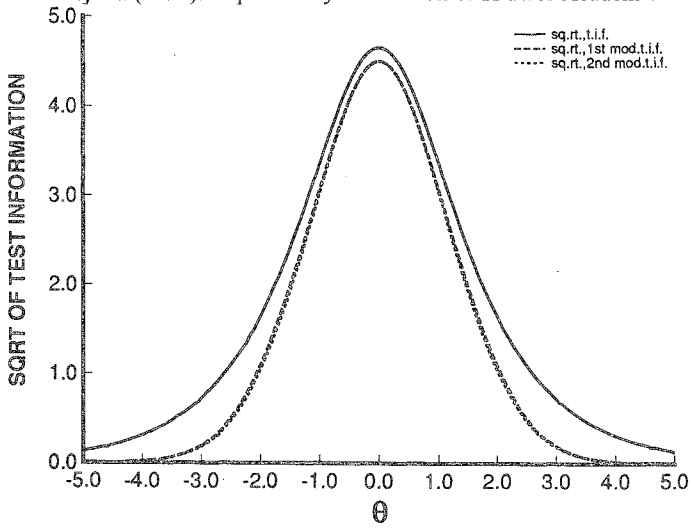
Table 1 shows that the predicted $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ obtained by using $I(\theta)$ was widely distributed; that is, it ranged from .20049 to .89641. $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ decreased as the mode of the distribution shifted from a range of θ in which the amount of test information was greater (e.g., Distributions 1 and 3) to another range in which it was lesser (e.g., Distribution 6). The reduction was more conspicuous when the SD of the normal distribution was smaller (compare Distributions 1 and 2 versus Distributions 3 and 4). The predicted $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$s obtained using $\Upsilon(\theta)$ indicated a substantial reduction from those using $I(\theta)$ for each of the six distributions of θ. For Distribution 2, $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ decreased from .82324 to .26479; for Distribution 5 from .47715 to .21681; and for Distribution 6 from .20049 to .01182. The reduction is more conspicuous for Distributions 2, 5, and 6, which were distributed on lower levels of θ where the discrepancies between $I(\theta)$ and $\Upsilon(\theta)$ were large. Among the six distributions, the predicted $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ obtained using $\Upsilon(\theta)$ varied from .01182 to .80074, showing even larger differences. Similar results were obtained for the predicted $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$s using $\Xi(\theta)$; these $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$s varied from .01109 to .79920. Also, for each distribution the reduction in $r\left(\hat{\theta}_1, \hat{\theta}_2\right)$ from that obtained by Equation 35 was relatively small, as expected from Figure 3. Similar results were obtained for the SEMs, in the reverse order, as shown in Table 2.

In CTT, the SEM, σ_E, is given by

$$\sigma_E = \left[\mathrm{Var}(X)\right]^{1/2}\left[1-r_{X_1X_2}\right]^{1/2}. \tag{41}$$

Careful observation of Table 1 reveals that there were substantial discrepancies between the values of σ_E

Figure 3
Square Roots of $I(\theta)$ (Solid Line), $\Upsilon(\theta)$ (Dashed Line), and $\Xi(\theta)$ (Dotted Line) for the Hypothetical 30-Item Test
[From Samejima (1994). Reprinted by Permission of Kluwer Academic Publishers]



obtainable by Equation 41 using the attributed $r(\hat{\theta}_1, \hat{\theta}_2)$s in Table 1 in place of $r_{X_1X_2}$ in Equation 41 and the corresponding SEMs, which were obtained by Equations 25, 39, and 40. For example, using the values of $r(\hat{\theta}_1, \hat{\theta}_2)$ obtained for Distribution 1 from the different TIF formulas [$I(\theta)$, $\Upsilon(\theta)$, and $\Xi(\theta)$], these results were .31914, .46453, and .47936, respectively; for Distribution 3 they were .21433, .22388, and .22475; and for Distribution 6 they were .44846, .49857, and .49876. There are differences in the degree that this set of three values differ from the corresponding set in Table 2. These different degrees of disagreement may be expected, for the degree of violation from the assumptions behind CTT was different for each distribution of θ.

The three predicted error variances of the MLE of θ are presented in Table 2, for each of the six hypothetical distributions. They were obtained using Equation 20, and by similar equations in which $I(\theta)$ was replaced by $\Upsilon(\theta)$ or $\Xi(\theta)$, respectively. θ was divided into small intervals of .05 width, and a number of rectangles were used for approximate integration. Simpson's quadrature formula (Elderton & Johnson, 1969) also could have been used and perhaps would have provided more accurate results.

Predicted Versus Empirical Reliabilities

Method. In order to evaluate the resulting predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s obtained by using $I(\theta)$, $\Upsilon(\theta)$, and $\Xi(\theta)$, a set of empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s for the six θ distributions were used as criteria based on simulated data. Follow-

Table 1
Obtained Mean and SDs of θ and Predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s for Each of the Six Distributions of θ Using $I(\theta)$, $\Upsilon(\theta)$, and $\Xi(\theta)$

θ Distribution	Mean of θ	SD of θ	$I(\theta)$	$\Upsilon(\theta)$	$\Xi(\theta)$
1	0.00000	.99157	.89641	.78053	.76629
2	−.80000	.99157	.82324	.26479	.25256
3	0.00000	.50155	.81738	.80074	.79920
4	−.80000	.50155	.73250	.66611	.65589
5	−1.60000	.50155	.47715	.21681	.20093
6	−2.40000	.50155	.20049	.01182	.01109

Table 2
Predicted SEMs and Theoretical Error SDs for Each of the Six θ Distributions Using $I(\theta)$, $\Upsilon(\theta)$, and $\Xi(\theta)$

θ Distribution	Predicted SEM			Theoretical Error SD		
	$I(\theta)$	$\Upsilon(\theta)$	$\Xi(\theta)$	$I(\theta)$	$\Upsilon(\theta)$	$\Xi(\theta)$
1	.30548	.37648	.38514	.11363	.27646	.29987
2	.37887	.64293	.66397	.21111	2.73003	2.90974
3	.23521	.24717	.24811	.05620	.06260	.06320
4	.29172	.32802	.33326	.09186	.12609	.13197
5	.48839	.73440	.76583	.27563	.90868	1.00034
6	.91974	2.76394	2.88922	1.00314	21.03633	22.43035

ing each of the six distributions of θ, a group of examinees was hypothesized. A response pattern of each hypothetical examinee was produced for the test and retest situations using the monte carlo method. Because the test consisted of 30 equivalent dichotomous test items, the number-correct (NC) test score was a sufficient statistic for the response pattern, and the MLE of θ was obtained as a one-to-one mapping of this sufficient statistic (Lord & Novick, 1968, p. 429). There were 1,998 hypothetical examinees for Distributions 1 and 2, and 2,004 for Distributions 3, 4, 5, and 6.

A problem arose as to how to deal with the $-\infty$s and $+\infty$s obtained as the MLEs of θ, before the correlation coefficient between the two sets of $\hat{\theta}_V$s could be computed. Table 3 presents the frequencies of these two extreme values in the test and retest situations separately, and of those in both situations, for each of the six distributions of θ. Although there were only three $-\infty$s in the retest and no other $-\infty$s or $+\infty$s for Distribution 3, more than half of the examinees in Distribution 6 obtained $-\infty$ as their MLE of θ in the initial test as well as in the retest, and more than one-third of the total examinees obtained $-\infty$ in both. As is common practice, the $-\infty$s and $+\infty$s were replaced by arbitrary single values of −2.65 and 2.65, respectively, and the correlations were computed using these values.

Table 3
Frequencies of $-\infty$s and $+\infty$s Obtained as MLEs of θ in the Initial Test and the Retest, and in Both, for Each of the Six Distributions of θ

θ Distribution	Test		Retest		Both		
	$-\infty$	$+\infty$	$-\infty$	$+\infty$	$-\infty$	$+\infty$	Total
1	56	47	43	49	19	20	1,998
2	197	4	195	6	90	3	1,998
3	0	0	3	0	0	0	2,004
4	56	0	45	0	14	0	2,004
5	437	0	399	0	161	0	2,004
6	1,143	0	1,118	0	732	0	2,004

Results. Test-retest reliability correlations, together with the two means, the two variances, and the covariance, are presented in Table 4. These results were compared to the predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s in Table 1. For Distribution 3, for which only three replaced values were used (see Table 4), the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s were very close to the predicted values; that is, .80724 versus .81738 [$I(\theta)$], .80074 [$\Upsilon(\theta)$], and .79920 [$\Xi(\theta)$]. Note, however, that in general the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$ became larger than the predicted value as the total frequency of the $-\infty$s and $+\infty$s increased. This enhancement is artificial, however, because by using −2.65 and 2.65 in place of $-\infty$ and $+\infty$, respectively, those who obtained $-\infty$ as their MLE of θ both in the test and retest situations were treated as if they obtained the same estimated θ in the two testing situations, even though the distance between the two $-\infty$s could be infinitely large, for example. The same logic applies for $+\infty$. Note that this enhancement did not occur for Distribution 4—.72334 versus .73250, .66611, and .65589

(see Tables 1 and 4). For Distribution 4, the frequencies of $+\infty$ were 0 in the test, the retest, and the combination of both, and the frequencies of $-\infty$ were as small as 56 and 45 in the test and retest situations, respectively, with only 14 individuals overlapping in both (i.e., the second smallest frequencies next to those for Distribution 3).

Table 4

Empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s for Each of the Six Distributions of θ Based on the MLEs of the Sample Examinees in the Test-Retest Situation Using Arbitrary Values of ±2.65 and ±2.12271 for Infinite Estimate Values and NC Scores, and Mean and Variance of ±2.65 Scores at Test and Retest and Their Covariances

θ Distribution	$r(\hat{\theta}_1, \hat{\theta}_2)$			Mean (±2.65)		Variance (±2.65)		Covariance
	±2.65	±2.12271	NC	Test	Retest	Test	Retest	(±2.65)
1	.90788	.91863	.93691	−.00311	.00106	1.19069	1.16769	1.07051
2	.88812	.90988	.93130	−.81435	−.80971	1.07982	1.09703	.96663
3	.80724	.80948	.82180	.00785	−.00754	.33578	.33443	.27051
4	.72334	.73948	.78857	−.85777	−.84349	.40504	.39310	.28863
5	.55304	.60701	.65453	−1.68722	−1.67511	.42299	.40820	.22980
6	.32187	.39757	.40075	−2.28115	−2.25897	.21639	.23189	.07210

The empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s also depended on the replacement values used for $-\infty$ and $+\infty$, especially when many examinees received $-\infty$ or $+\infty$ as their MLE of θ (compare Distributions 5 and 6 to Distributions 3 and 4). Because the two replacement values, −2.65 and 2.65, that were used in computing the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s presented in Table 4 were arbitrary, the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s were computed again by changing these two replacement values to −2.12271 and 2.12271, respectively. The results also are shown in Table 4. Comparing these values with those presented in the 2.65 column of Table 4, each of the values for the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s is greater than the corresponding value for 2.65. Although the increment is almost 0.0 for Distribution 3 and it is mild for Distribution 4, it is substantially large for Distributions 5 and 6. These results are predictable from the differences in the number of replacement values used for the different distributions (compare the frequencies of $-\infty$s and $+\infty$s for these distributions in Table 4).

The variability that exists among the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s across different distributions of θ might have been the result of using $\hat{\theta}_V$ rather than the NC score, although this interpretation is illogical. This is not the case, as is obvious from theory. To illustrate this, the empirical $r_{X_1X_2}$s also were computed using the NC for each distribution of θ. The results in Table 4 show the same type of variability in the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s across the six θ distributions as was observed when $\hat{\theta}_V$ was used. The empirical $r_{X_1X_2}$s based on NC were slightly higher than those obtained using $\hat{\theta}_V$ and 2.65 or 2.12271 in place of $-\infty$s and $+\infty$s. This results from the fact that 0 and 30 were used for the two indeterminant scores and artificially enhanced the correlations. A set of $r(\hat{\theta}_1, \hat{\theta}_2)$s based on $\hat{\theta}_V$ could be produced that are even closer to those based on the NC score by adjusting the replacement values for $-\infty$s and $+\infty$s.

When a subgroup of examinees who obtained the same NC score other than 0 and n is considered, it can be taken as an indicator of some sort of homogeneity among these in the subgroup. When the test score is either of these two extreme values, however, this type of homogeneity cannot be assumed, because giving 0 or n as the test score simply means that the test has failed in discriminating these examinees' θ levels. Because the $r_{X_1X_2}$ based on NC, or of the reliability coefficient based on $\hat{\theta}_V$ with an identical replacement value for each ∞, treats each of these two subgroups of examinees as if they had the same level of θ, the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s tend to be higher than they actually are, especially if there are many 0s or ns, or both. This explains the differences between the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s in Table 4 and the predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s in Table 1.

Estimating Upper Bounds of Empirical Reliabilities

It is obvious from the foregoing that an identical replacement value for each of the $-\infty$s and $+\infty$s should

not be used in obtaining the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s that are to be used as the criterion to evaluate the three predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s. Because there is no reasonable way to handle $-\infty$s and $+\infty$s, several attempts were focused on obtaining an estimate of the upper bound of the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$ for each θ distribution. Thus, in Method 1 the resulting $r(\hat{\theta}_1, \hat{\theta}_2)$s were obtained by:

1. Assigning a randomly produced true θ following the density function

$$\left[1 - P_g(\theta)\right]^n f(\theta)\left\{\int_{-\infty}^{\infty}\left[1 - P_g(\theta)\right]^n f(\theta)d\theta\right\}^{-1}, \tag{42}$$

   where n=30, to each examinee whose MLE of θ was $-\infty$. Then the amount of bias, $B(\theta; \hat{\theta}_V)$, was added to the θ thus assigned, and then the error score, produced by randomly following $N\{0, [I(\theta)]^{-1/2}\}$ was added, and the resulting value was substituted for the $-\infty$.
2. This same process was followed for each examinee whose MLE of θ was $+\infty$; however, the density function in Equation 42 was replaced by

$$\left[P_g(\theta)\right]^n f(\theta)\left\{\int_{-\infty}^{\infty}\left[P_g(\theta)\right]^n f(\theta)d\theta\right\}^{-1}. \tag{43}$$

   An identical value of true θ was assigned in the test and retest to an examinee who obtained $-\infty$ or $+\infty$ in both testing situations. The estimate of the upper bound of $r(\hat{\theta}_1, \hat{\theta}_2)$ for each of the six distributions of θ is presented in Table 5 (Method 1). Because of the way these substitute values for $-\infty$s and $+\infty$s were produced, it was expected that the results would be very conservative upper bounds for $r(\hat{\theta}_1, \hat{\theta}_2)$.

Table 5
Estimated Upper Bounds of $r(\hat{\theta}_1, \hat{\theta}_2)$ by Methods 1 and 2 for Each of the Six Distributions of θ Based on the MLEs of Sample Examinees in the Test-Retest Method

	$r(\hat{\theta}_1, \hat{\theta}_2)$			
θ Distribution	Method 1	Method 2	Method 1 Modified	Method 2 Modified
1	.84116	.89601	.82767	.82269
2	.71712	.89450	.58455	.63928
3	.81050	.81061	.81050	.81110
4	.72919	.72338	.72616	.71792
5	.41012	.56094	.34556	.35298
6	.36089	.61521	.12549	.12896

In Method 2, the results were obtained by using a truncated normal distribution to generate a true θ; that is, the population normal density function was divided by vertical lines into five segments whose areas were proportional to the frequencies of: (1) those who obtained $-\infty$ on both the test and retest, (2) those who obtained $-\infty$ in the test (or retest) situation only, (3) those who obtained finite values for their MLEs of θ in the test (or retest) situation, (4) those who obtained $+\infty$ in the test (or retest) situation only, and (5) those who obtained $+\infty$ on both the test and retest. The ordinate in each segment was divided by the total area of the segment to become densities. Again, an identical value of true θ was assigned in both the test and retest situations to an examinee who obtained $-\infty$ or $+\infty$ in both sessions. Because of the way in which these substitute values for $-\infty$s and $+\infty$s were produced, it was expected that the results would be even more conservative than those of Method 1.

The estimates of the upper bound of $r(\hat{\theta}_1, \hat{\theta}_2)$ for each of the six distributions of θ for Method 2 also are presented in Table 5. The estimates of the upper bounds of the $r(\hat{\theta}_1, \hat{\theta}_2)$s increased substantially over the results obtained by Method 1 when the data included a substantial number of $-\infty$ or $+\infty$s for which the

substitute values would be used (Distribution 6). This is because of how the true value of θ was assigned in Method 2; it made the variability in the substitute values for $-\infty$ too small.

Although the results of Method 1 provide legitimate upper bounds of the empirical $r(\hat{\theta}_1, \hat{\theta}_2)$s for the different θ distributions, it is obvious from theory that they are still far too conservative. Thus, procedures similar to the above two methods were followed without assigning an identical true θ to an examinee who obtained $-\infty$ or $+\infty$ in both the test and retest situations. The results also are shown in Table 5 (Method 1 Modified and Method 2 Modified). These two sets of results were similar to each other. Note that the two values for each distribution of θ were between the predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s obtained by Equations 21 [$I(\theta)$] and 35 [$\Upsilon(\theta)$], respectively (see Table 1).

Discussion and Practical Implications

The above examples were based on a hypothetical test of 30 equivalent items, which provided large discrepancies between the original TIF and the two modified TIF formulas. However, this test is not the kind of test usually used in practice. Figure 4 presents the square roots of the original TIF and its two modification formulas for two empirical tests—the Iowa Level 11 Vocabulary Subtest (Figure 4a), which consisted of 43 dichotomously scored items, and Shiba's Word/Phrase Comprehension Test J1 (Figure 4b), which consists of 54 dichotomously scored items (see Samejima, 1993a, 1993b).

Figures 4a and 4b show that (1) the three curves are flatter than those of the 30 equivalent test items, (2) the decrease in the amount of information is not as radical as θ departs from the modal point, and (3) the discrepancies between the square root of the original TIF and those of the modified formulas are not as conspicuous (Figure 4 versus Figure 3). However, the two curves for the modified formulas are almost overlapping, as was observed with the 30 equivalent test items (Figure 3).

Tables 1 and 5 showed that, although all three predictions of $r(\hat{\theta}_1, \hat{\theta}_2)$ were accurate when the distribution of θ was in the interval of θ in which the amount of test information was large, when the θ distribution shifted away from this interval Equation 35 or Equation 36, in which $I(\theta)$ in Equation 21 is replaced by $\Upsilon(\theta)$ or $\Xi(\theta)$, respectively, were better predictors than Equation 21. Thus the MLE bias function is useful in predicting $r(\hat{\theta}_1, \hat{\theta}_2)$. Considering that the values in Table 5 are upper bounds of the reliability coefficients, Equation 36 in which $\Xi(\theta)$ is used will be the most appropriate formula.

Thus the results of the present research suggest that it is advisable to use $\Xi(\theta)$ rather than $I(\theta)$ for predicting the reliability coefficient of a test attributed to a specific distribution of θ, as well as the criterion in the stopping rule of a CAT. With a finite number of test items the MLE of θ is conditionally biased, given θ. Therefore, in theory the use of $\Xi(\theta)$, which is based on the minimum bound of the mean squared error of $\hat{\theta}_v$ rather than the minimal variance bound (Samejima, 1990), is more reasonable than that of $\Upsilon(\theta)$ also, although the two results provided similar values here.

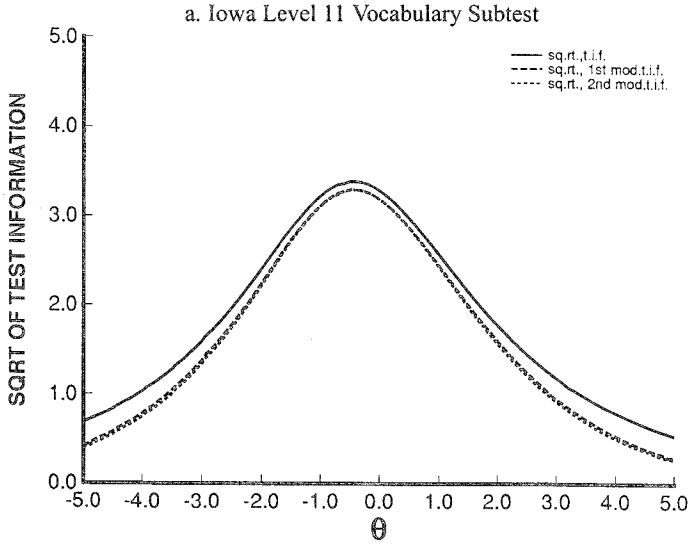
These examples were selected intentionally to make the differences among the different θ distributions and among the three predicted $r(\hat{\theta}_1, \hat{\theta}_2)$s for each θ distribution substantially large, using equivalent test items. Because equivalent test items are seldom used in actual tests, the differences between the resulting predicted reliability coefficients obtained by using $I(\theta)$ and by using either $\Upsilon(\theta)$ or $\Xi(\theta)$ are expected to be less.

Because of more useful and informative measures like the TIF and its two modified formulas, the reliability coefficient of a test is no longer important in modern test theory. It is interesting, however, to predict the reliability coefficient—which is attributed to separate groups of examinees—using the TIFs. The traditional concept of test reliability is misleading, because the reliability coefficient of the same test can be drastically different for different groups of examinees.

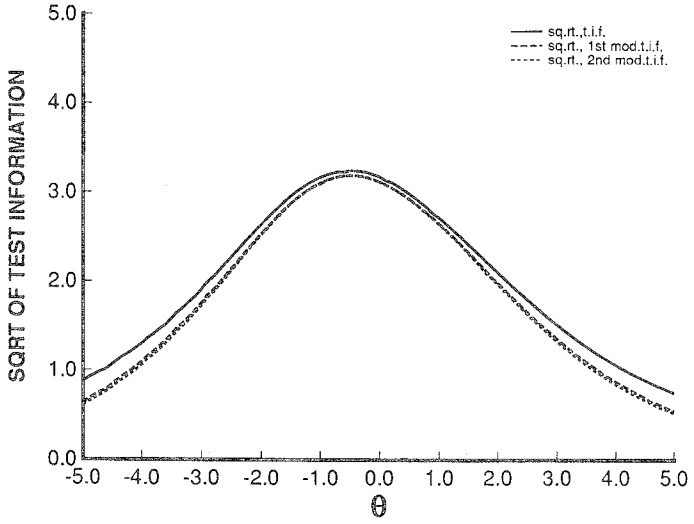
If the reliability coefficient must be used, a practical suggestion is to compute the predicted reliability coefficients attributed to as many different hypothetical distributions of θ as can be conceived for

Figure 4
Square Roots of $I(\theta)$ (Solid Line), $\Upsilon(\theta)$ (Dashed Line), and $\Xi(\theta)$ (Dotted Line) for Two Tests Following the Logistic Model

a. Iowa Level 11 Vocabulary Subtest



b. Shiba's Word/Phrase Comprehension Test J1



which the test is likely to be administered. In this way, one of the reliability coefficients can be selected, depending on the group from which the examinees being tested have been sampled. This will improve the current use of the reliability coefficient in which a single value is recorded in the test manual as *the* reliability coefficient and is used with no consideration of the population from which the sample of examinees have been selected.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

Elderton, W. P., & Johnson, N. L. (1969). *Systems of frequency curves.* London: Cambridge University Press.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61A,* 273–287.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* No. 7.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233–245.

Lord, F. M. (1984, April). *Technical problems arising in parameter estimation.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

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

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

Samejima, F. (1977a). Effects of individual optimization in setting boundaries of dichotomous items on accuracy of estimation. *Applied Psychological Measurement, 1,* 77–94.

Samejima, F. (1977b). A use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233–247.

Samejima, F. (1979). *Convergence of the conditional distribution of the maximum likelihood estimate, given latent trait, to the asymptotic normality: Observations made through the constant information model* (Office of Naval Research Rep. No. 79-3). Knoxville: Department of Psychology, University of Tennessee.

Samejima, F. (1981). *Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item* (Office of Naval Research Final Report of N00014-77-C-0360). Knoxville: Department of Psychology, University of Tennessee.

Samejima, F. (1987). *Bias function of the maximum likelihood estimate of ability for discrete item responses* (Office of Naval Research Report No. 87-1). Knoxville: Department of Psychology, University of Tennessee.

Samejima, F. (1990). *Modifications of the test information function* (Office of Naval Research Report No. 90-1). Knoxville: Department of Psychology, University of Tennessee.

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika, 58,* 119–138.

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58,* 195–209.

Samejima, F. (1994). Roles of Fisher type information in latent trait models. In H. Bozdogan (Ed.), *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach* (pp. 347–378). The Netherlands: Kluwer.

Acknowledgments

This research was supported by the Office of Naval Research Contracts N00014-87-K-0320 and N00014-90-J-1456.

Author's Address

Send requests for reprints or further information to Fumiko Samejima, 310B Austin Peay Bldg., University of Tennessee, Knoxville TN 37996-0900, U.S.A. Internet: samejima@psych1.psych.utk.edu.