PREDICTIONS OF RELIABILITY COEFFICIENTS AND STANDARD ERRORS OF MEASUREMENT USING THE TEST INFORMATION FUNCTION AND ITS MODIFICATIONS

FUMIKO SAMEJIMA

UNIVERSITY OF TENNESSEE
KNOXVILLE, TENN. 37996-0900

JULY, 1990

Prepared under the contract number N00014-87-K-0320, 4421-549 with the Cognitive Science Research Program Cognitive and Neural Sciences Division Office of Naval Research

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

R01-1069-11-005-91
Since we have more useful and informative measures like the test information function and its two modified formulae, the reliability coefficient of a test is no longer necessary in modern mental test theory. And yet it is interesting to know how to predict the coefficient using these functions, which are tailored for each separate population of examinees. In this process, it will become more obvious that the traditional concept of test reliability is misleading, for without changing the test the coefficient can be drastically different if we change the population of examinees.
TABLE OF CONTENTS

1 Introduction 1

2 Test Information Function and Its Modifications 1

3 Reliability Coefficient of a Test in the Sense of Classical Mental Test Theory 4
   3.1 General Case 4
   3.2 Maximum Likelihood Estimator 5

4 Standard Error of Measurement in the Sense of Classical Mental Test Theory 7

5 Examples 8

6 Discussion and Conclusions 17

REFERENCES
The research was conducted at the principal investigator's laboratory, 405 Austin Peay Bldg., Department of Psychology, University of Tennessee, Knoxville, Tennessee. Those who worked as assistants for this research include Christine A. Golik, Barbara A. Livingston, Lee Hai Gan and Nancy H. Domm.
I Introduction

There seems to be a consensus that two main measures in classical mental test theory are the reliability and validity coefficients of a test. Although these measures have widely been accepted by psychologists and test users in the past decades, they are actually the attributes of a specified group of examinees as well as of a given test, since the correlation coefficient is used in either case. In addition, representation of these measures by single numbers results in over-simplification and the lack of useful information for both theorists and actual users of tests. The same applies for the standard error of measurement also.

In latent trait models, the item and test information functions provide us with abundant information about the local accuracy of estimation, a concept which is totally missing in classical mental test theory. These functions do not depend upon any specific group of examinees as the reliability coefficient does, or we can say that they are population-free. By virtue of this characteristic, adding further information about the MLE bias function of the test and the ability distribution of the examinee group, we can provide the tailored reliability coefficient and standard error of measurement in the classical mental test theory's sense for each and every specified group of examinees who have taken the same test! (cf. Samejima, 1977b, 1987).

This progressive desolution of the reliability coefficient and of the standard error of measurement in classical mental test theory and their replacement by the test information function in latent trait models is further facilitated by the recent proposal of the modifications of the test information function, using the MLE bias function (cf. Samejima, 1987, 1990). In the present paper, it will be shown how we can predict the so-called reliability coefficient and standard error of measurement of a test in the sense of classical mental test theory, taking advantage of the new developments in latent trait models.

II Test Information Function and Its Modifications

Let \( \theta \) be ability, or latent trait, which takes on any real number. We assume that there is a set of \( n \) test items measuring \( \theta \) whose characteristics are known. Let \( g \) denote such an item, \( k_g \) be a discrete item 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 \( \theta \), i.e.,

\[
P_{k_g}(\theta) = \text{Prob.}[k_g \mid \theta].
\]

We assume that \( P_{k_g}(\theta) \) is three-times differentiable with respect to \( \theta \). We have for the item response information function (Samejima, 1972)

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

and the item information function, \( I_g(\theta) \), is defined as the conditional expectation of \( I_{k_g}(\theta) \), given \( \theta \), such that

\[
I_g(\theta) = E[I_{k_g}(\theta) \mid \theta] = \sum_{k_g} I_{k_g}(\theta) P_{k_g}(\theta).
\]

In the special case where the item \( g \) is scored dichotomously, this item information function is simplified to become
(2.4) \[ I'(\theta) = \left[ \frac{\partial}{\partial \theta} P_0(\theta) \right]^2 \left[ \{ P_0(\theta) \{ 1 - P_0(\theta) \} \}^{-1} \right], \]

where \( P_0(\theta) \) is the operating characteristic of the correct answer to item \( g \). Let \( V \) be a response pattern such that

(2.5) \[ V = \{ k_g \} \quad g = 1, 2, ..., n. \]

The operating characteristic, \( P_V(\theta) \), of the response pattern \( V \) is defined as the conditional probability of \( V \), given \( \theta \), and by virtue of local independence we can write

(2.6) \[ R_V(\theta) = \prod_{k_g \in V} P_k(\theta). \]

The response pattern information function (Samejima, 1972), \( I_V(\theta) \), is given by

(2.7) \[ I_V(\theta) = \frac{\partial^2}{\partial \theta^2} \log R_V(\theta) = \sum_{k_g \in V} I_k(\theta), \]

and the test information function, \( I(\theta) \), is defined as the conditional expectation of \( I_V(\theta) \), given \( \theta \), and we obtain from (2.2), (2.3), (2.5), (2.6) and (2.7)

(2.8) \[ I(\theta) = E[I_V(\theta) | \theta] = \sum_V I_V(\theta) R_V(\theta) = \sum_{\theta=1}^n I_\theta(\theta). \]

A big advantage of modern mental test theory is that the standard error of estimation can locally be defined by using \( \left[ I(\theta) \right]^{-1/2} \). Unlike its counterpart in classical mental test theory, this function does not depend upon the population of examinees, but is solely a property of the test itself, which should be the way if we call it the standard error, or the reliability, of a test. It is well known that this function provides us with the asymptotic standard deviation of the conditional distribution of the maximum likelihood estimate of \( \theta \), given its true value.

Lord has proposed a bias function for the maximum likelihood estimate of \( \theta \) in the three-parameter logistic model whose operating characteristic of the correct answer, \( P_0(\theta) \), is given by

(2.9) \[ P_0(\theta) = c_0 + (1 - c_0) [1 + \exp(-D a_0(\theta - b_0))]^{-1}, \]

where \( a_0 \), \( b_0 \), and \( c_0 \) are the item discrimination, difficulty, and guessing parameters, 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, which is denoted by \( B(\delta_V | \theta) \) in this paper, can be written as

(2.10) \[ B(\delta_V | \theta) = D[I(\theta)]^{-1} \sum_{g=1}^n a_g I_g(\theta)[\psi_g(\theta) - \frac{1}{2}], \]

where

(2.11) \[ \psi_g(\theta) = [1 + \exp(-D a_g(\theta - b_g))]^{-1}. \]
We can see in the above formula of Lord's MLE bias function that the bias should be negative when $\psi_0(\theta)$ is less than 0.5 for all the items, which is necessarily the case for some interval of $\theta$, $(-\infty, \theta_L)$, and should be positive when $\psi_0(\theta)$ is greater than 0.5 for all the items, which also necessarily happens for some interval, $(\theta_H, \infty)$, and in between the bias tends to be close to zero, for the last factor in this formula assumes negative values for some items and positive for some others, and, therefore, they cancel themselves out, provided that the difficulty parameter $b_d$ distributes widely.

Lord has applied this MLE bias function to an 85-item SAT Verbal test (Lord, 1984), and the result shows a fairly wide range of $\theta$ in which the bias is practically nil.

In the general case of discrete item responses, we obtain for the bias function of the maximum likelihood estimate (cf. Samejima, 1987)

$$B(\delta_\nu \mid \theta) = E[\delta_\nu - \theta \mid \theta] = -\frac{1}{2}[I(\theta)]^{-2} \sum_{g=1}^{n} \sum_{k_g} A_{k_g}(\theta) P''_{k_g}(\theta)$$

where $A_{k_g}(\theta)$ is the basic function 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 $\theta$, respectively. On the graded response level where item score $x_g$ assumes successive integers, 0 through $m_g$, each $k_g$ in the above formula must be replaced by the graded item score $x_g$. On the dichotomous response level, it can be reduced to the form

$$B(\delta_\nu \mid \theta) = -\frac{1}{2}[I(\theta)]^{-2} \sum_{g=1}^{n} P'_{k_g}(\theta) P''_{k_g}(\theta)[P_{k_g}(\theta)]^{-1} ,$$

with $P'_{g}(\theta)$ and $P''_{g}(\theta)$ indicating the first and second partial derivatives of $P_{g}(\theta)$ with respect to $\theta$, respectively. This formula includes Lord's bias function in the three-parameter logistic model as a special case.

Using this MLE bias function and taking the reciprocal of an approximate minimum variance bound of the maximum likelihood estimator, a modified test information function, $T(\theta)$, has been defined by

$$T(\theta) = I(\theta) \left[1 + \frac{\partial}{\partial \theta} B(\delta_\nu \mid \theta)\right]^{-2} ,$$

which is a reciprocal of an approximate minimum bound of the maximum likelihood estimator (cf. Samejima, 1990). From this formula, we can see that the relationship between this new function and the original test information function depends upon the first derivative of the MLE bias function. To be more precise, if the derivative is positive, then the new function will assume a lesser value than the original test information function. If it is negative, then this relationship will be reversed. If it is zero, i.e., if the MLE is conditionally unbiased, then these two functions will assume the same value.

The second modified test information function, $\Xi(\theta)$, is defined by

$$\Xi(\theta) = I(\theta) \left\{1 + \frac{\partial}{\partial \theta} B(\delta_\nu \mid \theta)\right\}^2 + I(\theta) \left[B(\delta_\nu \mid \theta)\right]^2 ,$$

which is the reciprocal of an approximate minimum bound of the mean squared error of the maximum likelihood estimator (cf. Samejima, 1990). We can see that the difference between the two modified test information functions, $T(\theta)$ and $\Xi(\theta)$, is the second and last term in the braces of the right hand side of formula (2.15). Since this term is nonnegative, we have
(2.16) \[ E(\theta) \leq T(\theta) \]

throughout the whole range of \( \theta \), regardless of the slope of the MLE bias function.

When the MLE bias function of the test is monotone increasing, as is the case with many existing tests, it is obvious from (2.14) that \( T(\theta) \) will assume no greater values than those of the original test information function \( I(\theta) \). The same applies to \( E(\theta) \), and we have the relationship,

\[ 2.17 \]
\[ E(\theta) \leq T(\theta) \leq I(\theta) \]

throughout the whole range of \( \theta \).

III Reliability Coefficient of a Test in the Sense of Classical Mental Test Theory

Although we can handle the concept of reliability much better in modern mental test theory by using the test information function, \( I(\theta) \), or one of its modification formulae, \( T(\theta) \) or \( E(\theta) \), it has also been observed (Samejima, 1977b) that, if we wish, the reliability coefficient of a test in the sense of classical mental test theory can be obtained easily from the observed data and the test information function under a general condition. Since we have two modification formulae of the test information function now, we are in a position that can handle the prediction of the reliability coefficient tailored for a specified population of examinees even better.

[III.1] General Case

Let \( \theta^*_v \) be any estimator of ability \( \theta \). We can write

\[ 3.1 \]
\[ \theta^*_v = \theta + \epsilon , \]

where \( \epsilon \) denotes the error variable. In the test-retest situation, we have

\[ 3.2 \]
\[ \begin{align*} 
\theta^*_v_1 &= \theta + \epsilon_1 \\
\theta^*_v_2 &= \theta + \epsilon_2,
\end{align*} \]

where the subscripts, 1 and 2, indicate the test and retest situations, respectively. If we can reasonably assume that in the test and retest situations:

\[ 3.3 \]
\[ Cov(\epsilon_1, \epsilon_2) = 0 , \]

\[ 3.4 \]
\[ Var(\epsilon_1) = Var(\epsilon_2) \]

and
(3.5) \[ \text{Cov}(\theta, e_1) = \text{Cov}(\theta, e_2) = 0 , \]
then we will have

(3.6) \[ \text{Corr}(\theta_{v_1}, \theta_{v_2}) = \frac{|\text{Var}(\theta_{v_1}) - \text{Var}(e_1)|}{\text{Var}(\theta_{v_1})} . \]

Note that if we replace ability \( \theta \) by one of its transformed forms, true test score \( T \), and use the observed test score \( X \) as the estimator of \( T \) and \( E \) as its error of estimation, then (3.1) can be rewritten in the form

(3.7) \[ T = X + E , \]
which represents the fundamental assumption in classical mental test theory, and (3.6) becomes a familiar formula for the reliability coefficient \( r_{xx} \),

(3.8) \[ r_{xx} = \frac{\text{Var}(T)}{\text{Var}(X)} . \]

In classical mental test theory, however, researchers seldom check if these assumptions are acceptable. In fact, in many cases (3.5) is violated if we replace \( \theta \) by \( T \), and \( e_1 \) and \( e_2 \) by \( E_1 \) and \( E_2 \), respectively, unless the test has been constructed in such a way that most individuals from the target population have mediocre true scores.

We can write in general

(3.9) \[ \text{Var}(e) = E[e - E(e)]^2 \]
\[ = E[e - E(e | \theta)]^2 + E[E(e | \theta) - E(e)]^2 \]
\[ + 2E[(e - E(e | \theta))(E(e | \theta) - E(e))] . \]

This indicates that, if the error variable \( e \) is conditionally unbiased for the interval of \( \theta \) of interest, then (3.9) will be reduced to the form

(3.10) \[ \text{Var}(e) = E[e^2] . \]

[III.2] **Maximum Likelihood Estimator**

Let \( \hat{\theta}_V \) or \( \hat{\theta} \) denote the maximum likelihood estimator of \( \theta \) based upon the response pattern \( V \). If 1) \( \hat{\theta} \) is conditionally unbiased for the interval of \( \theta \) of interest and 2) the test information function \( I(\theta) \) assumes reasonably high values for that interval, then we will be able to approximate the conditional distribution of \( \hat{\theta} \), given \( \theta \), by the normal distribution \( N(\theta, I(\theta)^{-1/2}) \) for the interval of \( \theta \) within which the examinees' ability practically distributes. Thus we have from (3.10)
When this is the case, from (3.6) we can write

\[ \text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = |\text{Var}(\hat{\theta}_1) - \mathbb{E}[\{I(\theta)\}^{-1}][\text{Var}(\hat{\theta}_1)]^{-1}]. \] (3.12)

Thus the reliability coefficient in the sense of classical mental test theory can be predicted by a single administration of the test, given the test information function \( I(\theta) \) and the ability distribution of the examinees.

It has also been observed that in computerized adaptive testing we can predict the reliability coefficient if a specified amount of test information is used for the stopping rule for a given level of ability in each of the test and retest situations, provided that the above two conditions 1) and 2) are met. In such a case, we can write

\[ \text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = |\text{Var}(\hat{\theta}_1) - \mathbb{E}[\{I(\theta)\}^{-1}][\text{Var}(\hat{\theta}_1)]^{-1}|
\frac{1}{2}, \] (3.13)

where \( I(\theta) \) and \( I(\theta) \) are the preset criterion test information functions in the test and retest situations, respectively, which are adopted as the stopping rules for the two separate situations. Note that these two criterion test information functions need not be the same, and also that the reliability coefficient is obtainable from a single administration. In a simplified case where, in each situation, the same amount of test information is used as the criterion for terminating the presentation of new items for every examinee, we can rewrite the above formula into the form

\[ \text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = |\text{Var}(\hat{\theta}_1) - \sigma_1^2][\text{Var}(\hat{\theta}_1)]^{-1}][\text{Var}(\hat{\theta}_1) - \sigma_2^2][\text{Var}(\hat{\theta}_1) + \sigma_2^2][\text{Var}(\hat{\theta}_1) + \sigma_2^2]]^{-1/2}, \] (3.14)

where \( \sigma_1^2 \) and \( \sigma_2^2 \) are the reciprocals of the constant amounts of criterion test information in the two separate situations, respectively. If we use the same constant amount of test information as the stopping rule in both the test and retest situations, then the reliability coefficient takes the simplest form

\[ \text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = |\text{Var}(\hat{\theta}_1) - \sigma^2][\text{Var}(\hat{\theta}_1)]^{-1}, \] (3.15)

where \( \sigma^2 \) 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 \( \theta \), can be examined by the Monte Carlo method (cf. Samejima, 1977a). We also notice that a necessary condition for this approximation is that \( \hat{\theta} \) is conditionally unbiased for the interval of \( \theta \) of interest. Thus we can use the MLE bias function, which was introduced in Section 2, for a test for the support of the approximation. Note that the MLE bias function together with the ability distribution of the target population also determines whether the assumption described by (3.5) should be accepted.
If the conditional unbiasedness is not supported, i.e., if \( B(\hat{\theta} \mid \theta) \) does not approximately equal zero for all values of \( \theta \) in the interval of interest, however, then we shall be able to adopt one of the modified test information functions, \( T(\theta) \) or \( \Xi(\theta) \). Thus we can rewrite (3.12) into the forms

\[
\text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = [\text{Var}(\hat{\theta}_1) - E[\{T(\theta)\}^{-1}]] \cdot \text{Var}(\hat{\theta}_1)^{-1}
\]

and

\[
\text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = [\text{Var}(\hat{\theta}_1) - E\{\Xi(\theta)\}^{-1}]] \cdot \text{Var}(\hat{\theta}_1)^{-1}.
\]

We can decide which of the modified formulae, (3.16) or (3.17), is more appropriate to use in a specified situation. Also in computerized adaptive testing, either \( T(\theta) \) or \( \Xi(\theta) \) can be used as the stopping rule in place of the test information function \( I(\theta) \), and we can revise (3.13) into the forms

\[
\text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = [\text{Var}(\hat{\theta}_1) - E[\{T(1)(\theta)\}^{-1}]] \cdot \text{Var}(\hat{\theta}_1)^{-1} \cdot [\text{Var}(\hat{\theta}_1) \cdot \text{Var}(\hat{\theta}_1) - E[\{T(1)(\theta)\}^{-1}]]^{-1/2},
\]

and

\[
\text{Corr}(\hat{\theta}_1, \hat{\theta}_2) = [\text{Var}(\hat{\theta}_1) - E\{\Xi(1)(\theta)\}^{-1}]] \cdot \text{Var}(\hat{\theta}_1)^{-1} \cdot [\text{Var}(\hat{\theta}_1) \cdot \text{Var}(\hat{\theta}_1) - E\{\Xi(1)(\theta)\}^{-1}]]^{-1/2},
\]

where the subscripts (1) and (2) represent the test and retest situations, respectively.

**IV Standard Error of Measurement of a Test in the Sense of Classical Mental Test Theory**

In classical mental test theory, the standard error of estimation of ability is represented by a single number, which is heavily affected by the degree of heterogeneity of the group of examinees tested, as is the case with the reliability coefficient. In contrast, in latent trait models, the standard error of estimation is *locally* defined, i.e., as a function of ability, which is the reciprocal of the square root of test information function. Since the test information function does not depend upon any specific group of examinees, but is a *sole* property of the test itself, this locally defined standard error is much more appropriate than the standard error of estimation in classical mental test theory. Also this function indicates that no test is efficient in ability measurement for the entire range of ability, and each test provides us with large amounts of information only locally, which makes a perfect sense to our knowledge.

The standard error of measurement of a test tailored for a specific ability distribution is given by

\[
S.E. = E[\{I(\theta)\}^{-1/2}]
\]

when the conditions 1) and 2) described in the preceding section are met, and by
S.E.1 = \mathrm{E}[(T(\theta))^{-1/2}]

or

S.E.2 = \mathrm{E}[(\Xi(\theta))^{-1/2}]

otherwise.

V Examples

For the purpose of illustration, six ability distributions are hypothesized, and for a single test predictions are made for their tailored reliability coefficients and tailored standard errors of measurement in the sense of classical mental test theory, using (3.12), (3.16), (3.17), (4.1), (4.2) and (4.3). These six hypothetical ability distributions are normal distributions, i.e., $N(0.0, 1.0)$, $N(-0.8, 1.0)$, $N(0.0, 0.5)$, $N(-0.8, 0.5)$, $N(-1.6, 0.5)$ and $N(-2.4, 0.5)$. Figure 5-1 presents the density functions of these six distributions. The hypothetical test consists of thirty equivalent dichotomous items, which follow the logistic model represented by (2.9) with $c_g = 0.0$, and the common parameter values $a_g = 1.0$ and $b_g = 0.0$, respectively, with the scaling factor $D$ set equal to 1.7. Figure 5-2 presents the MLE bias function of this hypothetical test. We can see in this figure that outside the interval of $\theta$, $(-1.0, 1.0)$, the amount of bias is substantially large. The square roots of the test information function $I(\theta)$ and of its two modification formulae $T(\theta)$ and $\Xi(\theta)$ of this test are shown in Figure 5-3.

Tables 5-1 and 5-2 present the resulting predicted reliability coefficients and standard errors of measurement for the six different ability distributions, respectively. In each table, the mean and the variance of $\theta$ of each of the six distributions are also given. We can see that these variances are slightly different from the squares of the second parameters of the normal distributions, i.e., 0.98322 vs. 1.00000 for the populations 1 and 2, and 0.25155 vs. 0.25000 for the populations 3, 4, 5 and 6, respectively, whereas all of the means are the same as the first parameters of the normal distributions. These discrepancies in variance come from the fact that we used frequencies for the equally spaced points of $\theta$ with the step width 0.05, which are given as integers, in order to approximate the normal distributions, instead of using the density functions themselves.

As you can see in the first table, the predicted reliability coefficient obtained by (3.12) distributes widely, i.e., it varies from 0.200 to 0.896! The coefficient reduces as the main part of the distribution shifts from a range of $\theta$ where the amount of test information is greater to another range where it is lesser. The reduction is more conspicuous when the standard deviation of the normal distribution is smaller. The predicted reliability coefficient obtained by (3.16) using $T(\theta)$ instead of $I(\theta)$ indicates a substantial reduction from the one obtained by (3.12) for each of the six ability distributions. The reduction is especially conspicuous for the populations 2, 5, and 6, whose ability distributions on lower levels of $\theta$ where the discrepancies between $I(\theta)$ and $T(\theta)$ are large. Among the six populations the predicted reliability coefficient obtained by means of (3.16) varies from 0.012 to 0.781, showing even a larger range than that obtained by (3.12). Similar results were obtained for the predicted reliability coefficient given by (3.17), using $\Xi(\theta)$ instead of $I(\theta)$. The reliability coefficient varies from 0.011 to 0.766, and within each population the reduction in the value of the reliability coefficient from the one obtained by (3.16) is relatively small, as is expected from Figure 5-3.

As for the standard error of measurement, we can see in Table 5-2 that similar results were obtained, only in reversed order, of course. In classical mental test theory, the standard error of measurement $\sigma_E$ is given by
FIGURE 5-1

Density Functions of Six Hypothetical Ability Distributions: \( n(0.0, 1.0) \), \( n(-0.8, 1.0) \), \( n(0.0, 0.5) \), \( n(-0.8, 0.5) \), \( n(-1.6, 0.5) \) and \( n(-2.4, 0.5) \).
MLE Bias Function of the Hypothetical Test of Thirty Equivalent Test Items Following the Logistic Model with \( a_g = 1.0 \) and \( b_g = 0.0 \) As the Common Parameters.
FIGURE 5-3

Square Roots of the Original (Solid Line) and the Two Modified (Dashed and Dotted Lines) Test Information Functions of the Hypothetical Test of Thirty Equivalent Items Following the Logistic Model with $a_g = 1.0$ and $b_g = 0.0$ as the Common Parameters.
Three Predicted Reliability Coefficients Tailored for Each of the Six Hypothetical Ability Distributions, Using the Original Test Information Function and Its Two Modification Formulae. The Indices, 1, 2 and 3, Represent the Original Test Information Function, Modification Formula No. 1 and Modification Formula No. 2, Respectively. The Mean and the Variance of $\theta$ for Each Population Are Also Given.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>RELIABILITY 1</th>
<th>RELIABILITY 2</th>
<th>RELIABILITY 3</th>
<th>MEAN OF THETA</th>
<th>VARIANCE OF THETA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.89641</td>
<td>0.78053</td>
<td>0.76629</td>
<td>0.00000</td>
<td>0.98322</td>
</tr>
<tr>
<td>2</td>
<td>0.82324</td>
<td>0.26479</td>
<td>0.25256</td>
<td>-0.80000</td>
<td>0.98322</td>
</tr>
<tr>
<td>3</td>
<td>0.81738</td>
<td>0.80074</td>
<td>0.79920</td>
<td>0.00000</td>
<td>0.25155</td>
</tr>
<tr>
<td>4</td>
<td>0.73250</td>
<td>0.66611</td>
<td>0.65589</td>
<td>-0.80000</td>
<td>0.25155</td>
</tr>
<tr>
<td>5</td>
<td>0.47715</td>
<td>0.21681</td>
<td>0.20093</td>
<td>-1.60000</td>
<td>0.25155</td>
</tr>
<tr>
<td>6</td>
<td>0.20049</td>
<td>0.01182</td>
<td>0.01109</td>
<td>-2.40000</td>
<td>0.25155</td>
</tr>
</tbody>
</table>
TABLE 5-2

Three Predicted Standard Errors of Measurement Tailored for Each of the Six Hypothetical Ability Distributions, Using the Original Test Information Function and Its Two Modification Formulae. The Indices, 1, 2 and 3, Represent the Original Test Information Function, Modification Formula No. 1 and Modification Formula No. 2, Respectively. The Mean and the Variance of $\theta$ for Each Population Are Also Given.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>STAND. ERROR 1</th>
<th>STAND. ERROR 2</th>
<th>STAND. ERROR 3</th>
<th>MEAN OF $\theta$</th>
<th>VARIANCE OF $\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.30548</td>
<td>0.37648</td>
<td>0.38514</td>
<td>0.00000</td>
<td>0.98322</td>
</tr>
<tr>
<td>2</td>
<td>0.37887</td>
<td>0.64293</td>
<td>0.66397</td>
<td>-0.80000</td>
<td>0.98322</td>
</tr>
<tr>
<td>3</td>
<td>0.23521</td>
<td>0.24717</td>
<td>0.24811</td>
<td>0.00000</td>
<td>0.25155</td>
</tr>
<tr>
<td>4</td>
<td>0.29172</td>
<td>0.32802</td>
<td>0.33326</td>
<td>-0.80000</td>
<td>0.25155</td>
</tr>
<tr>
<td>5</td>
<td>0.48839</td>
<td>0.73440</td>
<td>0.76583</td>
<td>-1.60000</td>
<td>0.25155</td>
</tr>
<tr>
<td>6</td>
<td>0.91974</td>
<td>2.76394</td>
<td>2.88922</td>
<td>-2.40000</td>
<td>0.25155</td>
</tr>
</tbody>
</table>
### TABLE 5-3

Three Theoretical Variances of the Maximum Likelihood Estimates of $\theta$ for Each of the Six Hypothetical Ability Distributions, Using the Original Test Information Function and Its Two Modification Formulae. The Indices, 1, 2 and 3, Represent the Original Test Information Function, Modification Formula No. 1 and Modification Formula No. 2, Respectively. The Mean and the Variance of $\theta$ for Each Population Are Also Given.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>VARIANCE OF MLE 1</th>
<th>VARIANCE OF MLE 2</th>
<th>VARIANCE OF MLE 3</th>
<th>MEAN OF THETA</th>
<th>VARIANCE OF THETA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.09684</td>
<td>1.25968</td>
<td>1.28308</td>
<td>0.00000</td>
<td>0.2322</td>
</tr>
<tr>
<td>2</td>
<td>1.19432</td>
<td>0.71324</td>
<td>3.89296</td>
<td>-0.80000</td>
<td>0.98322</td>
</tr>
<tr>
<td>3</td>
<td>0.30775</td>
<td>0.31414</td>
<td>0.31475</td>
<td>0.00000</td>
<td>0.25155</td>
</tr>
<tr>
<td>4</td>
<td>0.34341</td>
<td>0.37763</td>
<td>0.38352</td>
<td>-0.80000</td>
<td>0.25155</td>
</tr>
<tr>
<td>5</td>
<td>0.52718</td>
<td>1.16023</td>
<td>1.25189</td>
<td>-1.60000</td>
<td>0.5155</td>
</tr>
<tr>
<td>6</td>
<td>1.25469</td>
<td>21.28788</td>
<td>22.68190</td>
<td>-2.40000</td>
<td>5155</td>
</tr>
</tbody>
</table>
Three Theoretical Error Variances for Each of the Six Hypothetical Ability Distributions, Using the Original Test Information Function and Its Two Modification Formulae. The Indices, 1, 2 and 3, Represent the Original Test Information Function, Modification Formula No. 1 and Modification Formula No. 2, Respectively. The Mean and the Variance of $\theta$ for Each Population Are Also Given.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>VARIANCE OF ERROR 1</th>
<th>VARIANCE OF ERROR 2</th>
<th>VARIANCE OF ERROR 3</th>
<th>MEAN OF THETA</th>
<th>VARIANCE OF THETA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.11363</td>
<td>0.27646</td>
<td>0.29987</td>
<td>0.00000</td>
<td>0.98322</td>
</tr>
<tr>
<td>2</td>
<td>0.21111</td>
<td>2.73003</td>
<td>2.90974</td>
<td>-0.80000</td>
<td>0.98322</td>
</tr>
<tr>
<td>3</td>
<td>0.05620</td>
<td>0.06260</td>
<td>0.06320</td>
<td>0.00000</td>
<td>0.25155</td>
</tr>
<tr>
<td>4</td>
<td>0.09186</td>
<td>0.12609</td>
<td>0.13197</td>
<td>-0.80000</td>
<td>0.25155</td>
</tr>
<tr>
<td>5</td>
<td>0.27563</td>
<td>0.90868</td>
<td>1.00034</td>
<td>-1.60000</td>
<td>0.25155</td>
</tr>
<tr>
<td>6</td>
<td>1.00314</td>
<td>21.03633</td>
<td>22.43035</td>
<td>-2.40000</td>
<td>0.25155</td>
</tr>
</tbody>
</table>
Reliability Coefficient Computed for Each of the Six Hypothetical Ability Distributions Based upon the Maximum Likelihood Estimates of the Examinees for Test-Retest Situations Using a Test of Thirty Equivalent Items Following the Logistic Model with $D = 1.7$, $a_g = 1.0$ and $b_g = 0.0$. The Means and Variances of the Two Sessions and the Covariance Are Also Presented.

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>RELIABILITY</th>
<th>MEAN 1</th>
<th>MEAN 2</th>
<th>VARIANCE 1</th>
<th>VARIANCE 2</th>
<th>COVARIANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.90788</td>
<td>-0.00311</td>
<td>0.00106</td>
<td>1.19069</td>
<td>1.16769</td>
<td>1.07051</td>
</tr>
<tr>
<td>2</td>
<td>0.88812</td>
<td>-0.81435</td>
<td>-0.80971</td>
<td>1.07982</td>
<td>1.09703</td>
<td>0.96663</td>
</tr>
<tr>
<td>3</td>
<td>0.80724</td>
<td>0.00785</td>
<td>-0.00754</td>
<td>0.33578</td>
<td>0.33443</td>
<td>0.27051</td>
</tr>
<tr>
<td>4</td>
<td>0.72334</td>
<td>-0.85777</td>
<td>-0.84349</td>
<td>0.40504</td>
<td>0.39310</td>
<td>0.28863</td>
</tr>
<tr>
<td>5</td>
<td>0.55304</td>
<td>-1.68722</td>
<td>-1.67511</td>
<td>0.42299</td>
<td>0.40820</td>
<td>0.22980</td>
</tr>
<tr>
<td>6</td>
<td>0.32187</td>
<td>-2.28115</td>
<td>-2.25897</td>
<td>0.21639</td>
<td>0.23189</td>
<td>0.07210</td>
</tr>
</tbody>
</table>
\[
\sigma_E = [\text{Var}(X)]^{1/2} [1 - r_{XX}]^{1/2},
\]

where, as before, \( r_{XX} \) indicates the reliability coefficient. Comparison of Table 5-1 and Table 5-2 reveals that there are substantial discrepancies between the values of \( \sigma_E \) obtained by formula (5.1) using the tailored reliability coefficients in Table 5-1, which are based upon the maximum likelihood estimate \( \hat{\theta} \), in place of \( r_{XX} \), in (5.1) and the corresponding standard errors of measurement, which were obtained by formulæ (4.1) through (4.3) and presented in Table 5-2. To give some examples, for Population No. 1 the results of (5.1) are: 0.319, 0.465 and 0.479, respectively; for Population No. 3 they are: 0.214, 0.224 and 0.225; and for Population No. 6 they are: 0.448, 0.499 and 0.499. These results are understandable, for the degree of violation from the assumptions behind the classical mental test theory is different for the separate ability distributions.

The three theoretical variances of the maximum likelihood estimate of \( \theta \) and the three theoretical error variances are presented in Tables 5-3 and 5-4, respectively, for each of the six hypothetical populations. The latter were obtained by (3.11) and by replacing \( I(\theta) \) in (3.11) by \( T(\theta) \) and \( E(\theta) \), respectively, and the former are the sum of these separate error variances and the variance of \( \theta \).

In order to satisfy our curiosity, a simulation study has been made in such a way that, following each of the six ability distributions, a group of examinees is hypothesized, and using the Monte Carlo method a response pattern of each hypothetical subject is produced for each of the test and retest situations. Since our test consists of thirty equivalent dichotomous test items, the simple test score is a sufficient statistic for the response pattern, and the maximum likelihood estimate of \( \theta \) can be obtained upon this sufficient statistic. The numbers of hypothetical subjects are 1,998 for Populations No. 1 and No. 2, and 2,004 for Populations No. 3, No. 4, No. 5 and No. 6. The correlation coefficient between the two sets of \( \hat{\theta} \)'s was computed, and the results are presented in Table 5-5. Comparison of each of these results with the corresponding three tailored reliability coefficients in Table 5-1 gives the impression that, overall, these correlation coefficients are higher than the predicted tailored reliability coefficients. This enhancement comes from the fact that, in each distribution there are certain number of subjects who obtained negative or positive infinity as \( \hat{\theta} \), and we have replaced these negative and positive infinities by more or less arbitrary values, -2.65 and 2.65, respectively, in computing the correlation coefficients. Since in Population No. 3 none of the 2,004 hypothetical subjects got negative or positive infinity for their maximum likelihood estimates of \( \theta \) in the first session, and only three got negative infinity and none got positive infinity in the second session, this result, 0.807, will be the most trustworthy value. We can see that this value, 0.807, is less than 0.817 obtained by using the original test information function \( I(\theta) \), and a little greater than 0.801 obtained upon the Modification Formula No. 1, \( T(\theta) \). The next trustworthy value may be 0.723 of Population No. 4, for which none of the 2,004 subjects obtained positive infinity as their \( \hat{\theta} \)'s in each of the two sessions, and 56 and 45 got negative infinity in the first and second sessions, respectively. This value of correlation coefficient, 0.723, is a little less than the predicted reliability coefficient 0.733 obtained upon \( I(\theta) \), but somewhat greater than 0.666, which is based upon \( T(\theta) \), the Modification Formula No. 1—the artificial enhancement is already visible. The numbers of subjects who obtained negative and positive infinities in the first session and in the second session are: 56, 47, 43 and 49 for Population No. 1; 197, 4, 195 and 6 for Population No. 2; 437, 0, 399 and 0 for Population No. 5; and 1,143, 0, 1,118 and 0 for Population No. 6. We must say that, for these four distributions, the values of correlation coefficient in Table 5-5 should not be taken too seriously, for these values are enhanced because of the involvement of too many substitute values for negative and positive infinities.

VI Discussion and Conclusions

Test information function \( I(\theta) \) and its two modification formulæ, \( T(\theta) \) and \( E(\theta) \), are used to predict the reliability coefficient and the standard error of measurement which are tailored for each specific
ability distribution. Examples are given and a simulation study has been conducted for comparison.

These examples have been rather intentionally chosen to make the differences among the separate ability distributions, and among the three predicted indices for each ability distribution, clearly visible, using equivalent test items.

Since we have more useful and informative measures like the test information function and its two modified formulae, the reliability coefficient of a test is no longer necessary in modern mental test theory. And yet it is interesting to know how to predict the coefficient using these functions, which are tailored for each separate population of examinees. In this process, it will become more obvious that the traditional concept of test reliability is misleading, for without changing the test the coefficient can be drastically different if we change the population of examinees.

References

Distribution List

Dr. Terry Ackerman
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. James Algina
1403 Norman Hall
University of Florida
Gainesville, FL 32605

Dr. Erling B. Andersen
Department of Statistics
Studiestræde 6
1455 Copenhagen
DENMARK

Dr. Ronald Armstrong
Rutgers University
Graduate School of Management
Newark, NJ 07102

Dr. Eva L. Baker
UCLA Center for the Study
of Evaluation
145 Moore Hall
University of California
Los Angeles, CA 90024

Dr. Laura L. Barnes
College of Education
University of Toledo
2801 W. Bancroft Street
Toledo, OH 43606

Dr. William M. Bart
University of Minnesota
Dept. of Educ. Psychology
330 Burton Hall
178 Pillsbury Dr., S.E.
Minneapolis, MN 55455

Dr. Isaac Bejar
Mail Stop: 10-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Ramat Aviv 69978
ISRAEL

Dr. Arthur S. Blaives
Code N712
Naval Training Systems Center
Orlando, FL 32813-7100

Dr. Bruce Bloxom
Defense Manpower Data Center
Suite 155A
Monterey, CA 93941-3211

Cdt. Arnold Bohrer
Sectie Psychologisch Onderzoek
Rekruteringen- En Selectiecensuur
Kwartier Koningin Astrid
Bruijlnstraat
1120 Brussels, BELGIUM

Dr. Robert Breaux
Code 281
Naval Training Systems Center
Orlando, FL 32826-3224

Dr. Robert Brennan
American College Testing
Programs
P. O. Box 168
Iowa City, IA 52243

Dr. John B. Carroll
409 Elliott Rd., North
Chapel Hill, NC 27514

Dr. John M. Carroll
IBM Watson Research Center
User Interface Institute
P. O. Box 704
Yorktown Heights, NY 10598

Dr. Robert M. Carroll
Chief of Naval Operations
OP-01B2
Washington, DC 20350

Dr. Raymond E. Christal
UES LAMP Science Advisor
AFHR/LMB
Brooks AFB, TX 78235

Mr. Hua Hua Chung
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
Los Angeles, CA 90089-1061

Director, Manpower Program
Center for Naval Analyses
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Director,
Manpower Support and
Readiness Program
Center for Naval Analysis
2000 North Beauregard Street
Alexandria, VA 22311

Dr. Stanley Collyer
Office of Naval Technology
Code 222
800 N. Quincy Street
Arlington, VA 22217-5000

Dr. Hans F. Crombag
Faculty of Law
University of Limburg
P.O. Box 616
Maastricht
The NETHERLANDS 6200 MD

Ms. Carolyn R. Crane
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Dr. Timothy Davey
American College Testing Program
P.O. Box 168
Iowa City, IA 52243

Dr. C. M. Dayton
Department of Measurement
& Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Ralph J. DeAyala
Measurement, Statistics, and Evaluation
Benjamin Bldg., Rm. 4112
University of Maryland
College Park, MD 20742

Dr. Lou DiBello
CERL
University of Illinois
103 South Mathews Avenue
Urbana, IL 61801

Dr. Dattprasad Divgi
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. Hai-Ki Dong
Bell Communications Research
6 Corporate Place
PTA-1R226
Piscataway, NJ 08854

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Defense Technical
Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
(12 Copies)

5/1/90