DOCUMENT RESUME

ED 442 842 TM 031 262

AUTHOR	Samejima, Fumiko
TITLE	Some Considerations for Eliminating Biases in Ability Estimation in Computerized Adaptive Testing.
PUB DATE	1998-04-17
NOTE	35p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *Adaptive Testing; *Estimation (Mathematics); Item Response Theory; *Nonparametric Statistics; *Statistical Bias

ABSTRACT

Item response theory (IRT) has been adapted as the theoretical foundation of computerized adaptive testing (CAT) for several decades. In applying IRT to CAT, there are certain considerations that are essential, and yet tend to be neglected. These essential issues are addressed in this paper, and then several ways of eliminating noise and bias in estimating the individual parameter, theta, of person "a" are proposed and discussed, so that accuracy and efficiency in ability estimation can be increased. The content validity of the ability dimension is emphasized, and the idea of core test items is proposed. Devices are suggested to eliminate noise from multiple-choice items by using the nonparametric estimation of operating characteristics effectively in pilot studies. The use of the normal ogive model is suggested instead of the three-parameter logistic model. It is further suggested that several graded response items be used at the beginning of the CAT to avoid the influence of bias and lack of information inherent in dichotomous response items. The Weighted Likelihood Estimate of T. Warm (1989) and its expanded form for general discrete responses are discussed as an effective method of eliminating bias in ability estimation, and the usefulness of Warm's weight function as a prior is discussed. Use of the modified test information function is suggested for the same purpose. (Contains 10 figures and 18 references.) (SLD)

SOME CONSIDERATIONS FOR ELIMINATING BIASES IN ABILITY ESTIMATION IN COMPUTERIZED ADAPTIVE TESTING[1]

FUMIKO SAMEJIMA

UNIVERSITY OF TENNESSEE

The 1998 Annual AERA Meeting

April 17, 1998

San Diego, California

[1] Requests for reprints should be sent to Fumiko Samejima, Department of Psychology, 405 Austin Peay Bldg., University of Tennessee, Knoxville, Tennessee 37996-0900. E-mail: samejima@psych1.psych.utk.edu

Item response theory (IRT) has been adopted as the theoretical foundation of computerized adaptive testing (CAT) for several decades. In applying IRT for CAT, however, there are certain considerations that are essential, and yet tend to be neglected. In this paper, first these essential issues are addressed and discussed, and then several ways of eliminating noise and bias in estimating the individual parameter θ_a of person a are proposed and discussed, so that accuracy and efficiency in ability estimation be increased.

I. CONTENT VALIDITY OF THE ABILITY DIMENSION

[I.1] Necessity of Operational Definition of Ability θ

There has been a tendency that, once methodologies have been developed in IRT and accomodated in computer software, researchers apply them rather mechanically, without questioning if their target of estimation, ability θ, is properly defined in the process. Without due considerations for this issue, however, all our effort will be meaningless, and we will end up with obtaining mere artifacts that are of little psychological and educational significance. To give a concrete example, there is no guarantee that the θ's measured by LOGIST and BILOG are the same ability even if they are based on the same set of data, and yet very few researchers raise this question.

Thus an operational definition of θ is by far the most important in applying IRT for educational and psychological data. Although ability θ tends to be simply assumed, and its unidimensionality is taken for granted, we must start with defining θ operationally, and confirm its unidimensionality.

[I.2] Mathematical Challenge and Contribution to Education

In developing theories based on mathematics, there usually is a great deal of mathematical challenge that motivates psychometricians to work on specific topics. Thus we owe valuable outcomes in IRT to those theorists who have accepted such a challenge, conquered difficult problems and provided us with methodologies.

Too much emphasis on mathematical challenge sometimes makes us lose perspective, however. Take an example in simultaneous estimation of the individual parameter θ_a of a person a and the item parameters following some mathematical model. There is no doubt that this topic involves a great deal of mathematical challenge, and yet we must wonder if it is legitimate to estimate both individual and item parameters simultaneously.

It is advisable to keep in mind that our objective is to estimate the individual parameter θ_a , and that test items are only tools with which θ_a is estimated. Thus whenever necessity arises we can change or replace those *human-made* test items. In defining ability θ operationally, a set of items that reflects the target ability must be carefully selected so that the content validity of the resulting ability dimension be assured.

[I.3] Core Test Items

Suppose we have a set of test items whose content validity are assured from our past research findings. Let us call them *core test items.* If we succeed in extracting a single pricipal common factor behind these items, then we may accept it as the operationally defined θ . If we do not, then factor structure of those common factors should be examined, and appropriate deletion and/or addition of some items will eliminate minor clusters to provide a single principal common factor.

Ability θ thus operationally defined should have content validity, and will be used for item calibration of all items in the itempool. This is especially useful in on-line item calibration. Note that those core items do not have to be included in the itempool. To give an example, suppose, for practicality, we need to use only dichotomous response items in CAT. We can still include graded response items in the set of core test items, and in fact it is desirable to do so because:

1. in general, the amount of item information provided by a graded response item is greater than that of a dichotomous response item (see Samejima, 1969), and
2. more logical reasoning processes can be accomodated in a graded response item than in a dichotomous response item.

Insert Figure 1 About Here

Figure 1 presents a set of example questions taken from LSAT, the Official Prep Test III, 1991, Vol. 2. We could make a single graded response item out of these questions with the grades $0, 1, 2, 3, 4$, as illustrated in Figure 2. In this example, score 1 is given to those who found out the positions of J and T directly from the statements (e) and (f), score 2 is given to those who discovered, in addition, indeterminancy of the positions of K and L and that of X , Y and Z from the statements (b), (c) and (d), score 3 to those who found out the position of U , and score 4 to those who discovered the positions of K and L in the *if* situation given by the statement (g). This type of graded response item will be appropriate for a core test item because of its abundant item information for a wide range of θ and the fact that it represents logical reasoning processes necessary for grasping both what we can say and what we cannot say based on the statements. Note that we could increase the number of grade categories to 6, 7 or more if we further elaborate *if questions* exemplified by (g).

Insert Figure 2 About Here

II. ELIMINATION OF NOISE CAUSED BY GUESSING

[II.1] Unique Maximum Condition

Let g $(= 1, 2, \ldots, n)$ denote an item, which elicits any discrete response. Let $P_{k_g}(\theta)$ be the operating characteristic of the discrete response $K_g = k_g$ defined by

$$P_{k_g}(\theta) \equiv prob.\, [K_g = k_g \mid \theta] \ , \tag{1}$$

with the assumption that $P_{k_g}(\theta)$ is, at least, five times differentiable with respect to θ .

Samejima (1969, 1972) defined the basic function $A_{k_g}(\theta)$ such that

$$A_{k_g}(\theta) = \frac{\partial}{\partial\theta} \log P_{k_g}(\theta) \ . \tag{2}$$

Samejima (1973) also defined the item response information function, $I_{k_g}(\theta)$, which is given by

$$I_{k_g}(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_{k_g}(\theta) \ , \tag{3}$$

and the item information function $I_g(\theta)$ is obtained as the conditional expectation, given θ , of the item response information function, that is,

$$I_g(\theta) = E[I_{k_g}(\theta) \mid \theta] = \sum_{k_g} I_{k_g}(\theta)\ P_{k_g}(\theta) \ . \tag{4}$$

Eq. (4) includes Birnbaum's (1968) item information function for a dichotomous item as a special case.

The response pattern V is given by

$$V' = (K_1,\ K_2,\ K_3,\ \ldots\ldots,\ K_n) \ , \tag{5}$$

and due to local independence (Lord & Novick, 1968) the likelihood function $L(v \mid \theta)$ for general discrete responses can be written as

$$L(v \mid \theta) = P_v(\theta) \equiv prob.[V = v \mid \theta] = \prod_{k_g \in v} P_{k_g}(\theta) \ , \tag{6}$$

where $P_v(\theta)$ is the operating characteristic of the response pattern $V = v$. From Eqs. (3) and (6) the response pattern information function, $I_v(\theta)$, is given by

$$I_v(\theta) \;=\; -\frac{\partial^2}{\partial\theta^2}\,\log P_v(\theta) \;=\; \sum_{k_g \in v} I_{k_g}(\theta) \;\; . \tag{7}$$

The test information function, $I(\theta)$, is defined as the conditional expectation of the response pattern information function, given θ , and from Eqs. (3), (4), (6) and (7) we obtain

$$I(\theta) = E[I_v(\theta) \mid \theta] \;=\; \sum_{v} I_v(\theta)\; P_v(\theta) = \sum_{g=1}^{n} I_g(\theta) \;\; . \tag{8}$$

It is obvious from Eqs. (2) and (6) that for a test of n graded response items there are only $\sum_{g=1}^{n} m_g + n$ basic functions defined by Eq. (2). Using this small number of basic functions, a simple algorithm provides $\prod_{g=1}^{n}(m_g + 1)$ likelihood equations and hence the same numbers of maximum likelihood estimates (MLE's) $\hat{\theta}_v$'s . For example, if $n = 10$ and $m_g = 2$ for all items, then 30 basic functions provide MLE $\hat{\theta}_v$'s for as many as $59,049$ different response patterns. When all items are scored dichotomously, the number of basic functions is $2n$ and they provide 2^n MLE's.

Samejima (1969, 1972) proposed a sufficient condition for a discrete item response to provide a unique local or terminal maximum likelihood estimate for every response pattern consisting of such item responses. The condition is that the basic function $A_{k_g}(\theta)$, defined by Eq. (2), be strictly decreasing in θ with non-negative and non-positive values for its two asymptotes, respectively. For brevity, this condition has often been referred to as the *unique maximum condition.* It is noted from Eqs. (2) and (3) that the first part of this condition can be rephrased, that is, the item response information function $I_{k_g}(\theta)$ be positive for all θ except, at most, at an enumerable number of points where it may assume zero.

It has been shown (Samejima, 1969, 1972) that the unique maximum condition is satisfied by both the normal ogive model and the logistic model for dichotomous responses. Let $P_g(\theta)$ be the item characteristic curve (ICC), which is defined by

$$P_g(\theta) \;\equiv\; prob.\,[U_g = 1 \mid \theta] \;\; ,$$

where U_g $(= 0, 1)$ is a binary item score of item g with u_g as its realization. In the normal ogive and logistic models, ICC's are given by

$$P_g(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_g(\theta - b_g)} \exp[-\frac{u^2}{2}]\, du \ , \tag{9}$$

and

$$P_g(\theta) = \frac{1}{1 + \exp\{-Da_g(\theta - b_g)\}} \ , \tag{10}$$

respectively, where a_g (> 0) and b_g are the discrimination and difficulty parameters and $D = 1.702$ in Eq. (10) is a scaling factor.

It has also been shown (Samejima, 1972) that both the normal ogive and logistic models for graded responses satisfy the unique maximum condition, and so does Bock's nominal response model (Bock, 1972), which includes both Masters' partial credit model (Masters, 1982) and Muraki's generalized partial credit model (Muraki, 1992) for graded responses as special cases (see Samejima, 1972). It has also been proved that all models that belong to the logistic positive exponent family (Samejima, 1997) satisfy the same condition. Thus in these models the likelihood function that is based on the response pattern has a unique local or terminal maximum for every $v \in V$.

It should be noted, however, that the three-parameter logistic model (3PL), whose ICC is given by

$$P_g(\theta) = c_g + (1 - c_g)\Psi_g(\theta) \ , \tag{11}$$

where

$$\Psi_g(\theta) = \frac{1}{1 + \exp\{-Da_g(\theta - b_g)\}}$$

and c_g is the third parameter called the guessing parameter, does not satisfy the unique maximum condition (Samejima, 1972, 1973), and thus for some response patterns the likelihood functions may have multi-modes. Yen, Burket & Sykes (1991) have shown that multi-modality of the likelihood function occurs not infrequently for response patterns that usually come across in empirical data when the 3PL is used.

[II.2] Suggestion of the Use of the Normal Ogive Model

It is a common practice in CAT that the 3PL is adopted as the mathematical model for multiple-choice test items in the itempool. Since the third parameter c_g in Eq. (11) is nothing but noise that lowers the accuracy of estimation of the individual parameter θ_a , as is obvious from the fact that the 3PL does not even satisfy the unique maximum condition, it is desirable to replace it with some other model that includes less noise and, therefore, provides greater accuracy in ability estimation.

To realize this, first of all we must develop test items whose ICC's do not include so much noise within the framework of multiple choice format. Samejima (1994a) distinguished informative distractors from equivalent distractors, and called the operating characteristic of an informative distractor the plausibility function. Suppose we have developed an item whose distractors have differential information, in the sense that they tend to attract examinees of different levels of ability. In practice, it is desirable to include a distractor whose plausibility is identified by examinees of substantially high levels of ability, another distractor which attracts examinees of slightly lower levels of ability, etc., down to a distractor which attracts examinees of very low levels of ability. In the noiseless situation we can treat such an item as a graded response item.

Insert Figure 3 About Here

Figure 3 illustrates the ICC of such an item by a solid line, and the plausibility functions of the 4 distractors by dashed lines of various lengths in the noiseless situation, following the normal ogive model for graded responses (Samejima, 1969, 1972). In this model, the operating characteristics, $P_{x_g}(\theta)$'s , are given by

$$P_{x_g}(\theta) = P^*_{x_g}(\theta) - P^*_{(x_g+1)}(\theta) \ , \tag{12}$$

where $P^*_{x_g}(\theta)$ is called the cummulative operating characteristic of the graded item score $x_g\ (= 0, 1, 2, ..., m_g)$ of item g which is given by

$$P^*_{x_g}(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_g(\theta - b_{x_g})} \exp[-\frac{u^2}{2}]\, du \quad , \qquad (13)$$

where $a_g\ (> 0)$ is the item discrimination parameter, and b_{x_g} is the item response difficulty parameter which satisfies

$$-\infty = b_0 < b_1 < ... < b_{m_g} < b_{m_g+1} = \infty \ .$$

In the present example, the parameters in Eq. (13) are $a_g = 1.0$ and $b_{x_g} = -1.50, -0.50, 0.00, 0.75, 1.25$ respectively.

In the noiseless situation no guessing occurs, and examinees who do not find plausibility in any of the 5 alternative answers are supposed to honestly check the additional category, *don't know.* The strictly decreasing curve with the longest dashes in Figure 3 represents the operating characteristic of this don't-know category.

In practice, however, we cannot expect such total honesty, and it is likely that examinees in the don't-know category turn to guessing. Figure 4 presents the operating characteristics of the five alternative answers assuming that examinees in the don't-know group guess randomly.

Insert Figure 4 About Here

Figure 5 presents the ICC taken from Figure 4 in comparison with the one following the normal ogive model for dichotomous responses whose ICC is given by Eq. (9) with $a_g = 1.00$ and $b_g = 1.25$, together with the ICC in the 3PL with $c_g = 0.2$ that fits the ICC in question very well except on lower levels of θ . If we accept the ICC in the 3PL for this item, the critical value $\underline{\theta}_g$ equals 1.096 , and below this value the uniqueness of the MLE is not assured (see Samejima, 1973). The item is, therefore, not appropriate to use for examinees whose individual parameters θ_a 's are below this value of θ .

Insert Figure 5 About Here

If we accept the ICC in the normal ogive model as an approximation, however, the fit is extremely good for the interval of θ , $(0.3, \infty)$. Since it is not likely that, in CAT, this item is used for examinees whose θ_a 's are lower than 0.3 at which the ICC is as low as 0.15 , in practice this item can be treated as a *noise-free* item, and the use of normal ogive model will be justified.

Birnbaum (1968) proposed the logistic model whose ICC is given by Eq. (10) as a substitute for the normal ogive model. A strength of the logistic model lies in its mathematical simplicity, that includes a sufficient statistic $\sum_{g=1}^{n} u_g \; a_g$ which enables us to obtain the MLE without even using a computer. In this day of advanced computer technologies, however, we can use the normal ogive model just as easily, so there is no need for any substitute models.

[II.3] Effective Use of Nonparametric Approach

In order to find out if the distractors of our item are informative or not, we must *discover* their plausibility functions. For this purpose, nonparametric approaches for the estimation of operating characteristics, which do not a priori assume any mathematical forms, are by far the most useful.

Insert Figure 6 About Here

Figure 6 exemplifies the results obtained by using the simple sum procedure of the conditional p.d.f. approach (see Samejima, 1998) for the multiple-choice items of the Iowa Level 11 Vocabulary Subtest. Thus it has been disclosed that the item represented by the upper graph has informative distractors, while the distractors of the item represented by the lower graph do not provide differential information, and, therefore, are equivalent distractors. Because test items are human-made, if in pilot studies we discover that their disctractors belong to the second category, we can replace them by more informative ones that belong to the first category. Such pilot studies can be conducted in on-line item calibration, which is appropriate for CAT.

III. INITIAL TEST ITEMS IN CAT

Figure 7 presents the square roots of the test information function, which is given by Eq. (8), of hypothetical 30 equivalent dichotomous test items following the normal ogive model, with the common difficulty parameter $b_g = 0$ for all items of the five tests and the common discrimination parameters $a_g = 0.4, 0.7, 1.0, 1.5, 2.0$ for items of the separate tests, respectively. This square root of the test information function, $\sqrt{I(\theta)}$, is the reciprocal of the asymptotic standard error of estimation specified as a function of θ .

Insert Figure 7 About Here

Figure 7 implies that, although the minimal estimation error is smaller when the common item discrimination parameter is larger, the interval of θ for which the error is sufficiently small is narrower. Thus contrary to the general belief the use of items with high discrimination parameters may not be desirable, especially at the initial stage of CAT where the examinee's estimated individual parameter fluctuates for a relatively wide range.

This is also supported by the fact that, if items have higher discrimination parameters, the MLE bias function, $B(\theta; \hat{\theta}_v)$, which is given by

$$B(\theta; \hat{\theta}_v) = -\frac{1}{2[I(\theta)]^2} \sum_{g=1}^{n} \frac{\frac{\partial}{\partial\theta} P_g(\theta) \frac{\partial^2}{\partial\theta^2} P_g(\theta)}{P_g(\theta) Q_g(\theta)} \quad , \tag{14}$$

for a test of dichotomous items in general where

$$Q_g(\theta) \ = \ 1 - P_g(\theta) \ \ , \tag{15}$$

has a narrower interval of θ for which the MLE is practically unbiased. This is illustrated in Figure 8 for the same five hypothetical tests of 30 equivalent dichotomous items.

Insert Figure 8 About Here

A better solution for this problem than the use of low discrimination items at the initial stage of CAT may be the use of several graded response items such as the one illustrated in [I.3]. Since in general a graded response item provides a greater amount of information, and also a wider interval of θ for which MLE is practically unbiased, than a dichotomous item, its use will be an ideal solution.

IV. ELIMINATION OF BIAS IN ABILITY ESTIMATION

[IV. 1] Warm's Weighted Likelihood Estimate

A class of Bayesian modal estimators, θ_v^*, of ability θ can be defined as the value of θ that maximizes

$$L(v \mid \theta)\ f(\theta) \quad ,$$

where $L(v \mid \theta)$ is the likelihood function of a specific response pattern $V = v$, and $f(\theta)$ is known as a *prior*. Thus θ_v^* is the solution of

$$\frac{\partial}{\partial\theta}\log L(v \mid \theta) + \frac{\partial}{\partial\theta} f(\theta) \equiv 0 \quad .$$

When all the n items are scored dichotomously, the response pattern V takes the form of

$$V' = (U_1,\ U_2,\ U_3,\ ......,\ U_n) \quad .$$

By local independence the likelihood function $L(v \mid \theta)$ can be written as

$$L(v \mid \theta) = \prod_{g=1}^{n} P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \quad , \tag{16}$$

where $Q_g(\theta)$ is given by Eq. (15).

Lord (1983) proposed the bias function of the MLE, which is denoted by $B(\theta;\hat{\theta}_v)$ in this paper, for 3PL in which the ICC is given by Eq. (11). This bias function is given by

$$B(\theta;\hat{\theta}_v) \;=\; D[I(\theta)]^{-2}\sum_{g=1}^{n} a_g I_g(\theta)[\Psi_g(\theta) - \frac{1}{2}] \ . \tag{17}$$

Warm's (1989) weighted likelihood estimate (WLE) was proposed in the effort of minimizing the bias of θ_v^* by setting an appropriate prior, which he denoted $w(\theta)$. This prior can be expressed by the equation

$$\frac{\partial}{\partial\theta}\log w(\theta) \;=\; -B(\theta;\hat{\theta}_v)\ I(\theta) \ \ ,$$

where $B(\theta;\hat{\theta}_v)$ is the MLE bias function in 3PL given by Eq. (17), and $I(\theta)$ is the test information function that can be written as

$$I(\theta) \;=\; \sum_{g=1}^{n} \frac{\frac{\partial}{\partial\theta}P_g(\theta)}{P_g(\theta)\ Q_g(\theta)} \tag{18}$$

for general dichotomous responses (Birnbaum, 1968). Thus the WLE, which is denoted by $\tilde{\theta}_v$ in this paper, is the solution of:

$$\frac{\partial}{\partial\theta}\log L(v \mid \theta) + \frac{\partial}{\partial\theta}w(\theta) \;=\; \sum_{g=1}^{n} \frac{[u_g - P_g(\theta)]\ \frac{\partial}{\partial\theta}P_g(\theta)}{P_g(\theta)Q_g(\theta)} \;-\; B(\theta;\hat{\theta}_v)I(\theta) \;\equiv\; 0 \ . \tag{19}$$

[IV. 2] Expansion of the WLE for General Discrete Responses

Samejima (1993a, 1993b) expanded Lord's MLE bias function in 3PL for any discrete responses K_g 's , for which the response pattern V is given by Eq. (5). This MLE bias function for general discrete responses is given by

$$B(\theta;\hat{\theta}_v) \;=\; -\frac{1}{2[I(\theta)]^2}\sum_{g=1}^{n}\sum_{k_g} \frac{\frac{\partial}{\partial\theta}P_{k_g}(\theta)\ \frac{\partial^2}{\partial\theta^2}P_{k_g}(\theta)}{P_{k_g}(\theta)} \ \ , \tag{20}$$

where $P_{k_g}(\theta)$ and $I(\theta)$ are defined by Eqs. (1) and (8), respectively. When K_g is replaced by the graded item score X_g , all k_g 's in Eq. (20) are changed to x_g 's $(= 0, 1, 2, ..., m_g)$; when it is replaced by the binary item score U_g , Eq. (20) becomes Eq. (14) which includes Eq. (17) as a special case.

A straight-forward expansion of Warm's WLE for 3PL to general discrete responses provides the solution of:

$$\frac{\partial}{\partial\theta} \log\ L(v \mid \theta) + \frac{\partial}{\partial\theta} w(\theta) = \sum_{k_g \in v} A_{k_g}(\theta) - B(\theta; \hat{\theta}_v) I(\theta) \equiv 0 \tag{21}$$

as the WLE $\tilde{\theta}_v$, where $L(v \mid \theta)$ is given by Eq. (6), $A_{k_g}(\theta)$ is the basic function defined by Eq. (2), $B(\theta;\hat{\theta}_v)$ is the MLE bias function given by Eq. (20), and $I(\theta)$ is the test information function for general discrete responses provided by Eq. (8).

[IV. 3] Graphical Comparison of the WLE with the MLE

A graphical representation of the MLE and Warm's WLE will make their comparison easy. Figure 9 presents the MLE bias function $B(\theta;\hat{\theta}_v)$ of a hypothetical test of 30 equivalent dichotomous items following the normal ogive model, whose ICC is given by Eq. (9), with the common discrimination parameter $a_g = 0.7$ and the common difficulty parameter $b_g = 0.0$, respectively, represented by a short dashed line, and its product with the test information function $I(\theta)$ by a long dashed line. In the same figure, also presented are $\frac{\partial}{\partial\theta}$ log $L(v \mid \theta)$ for four response patterns, which include 0, 1, 7, and 15 correct answers, respectively.

Insert Figure 9 About Here

It is obvious from Eq. (21) that the WLE $\tilde{\theta}_v$ of each response pattern is the value of θ at which $\frac{\partial}{\partial\theta}$ log $L(v \mid \theta)$ crosses the long dashed curve representing $B(\theta;\hat{\theta}_v)\ I(\theta)$, while the MLE $\hat{\theta}_v$ of the same response pattern is that of θ at which $\frac{\partial}{\partial\theta}$ log $L(v \mid \theta)$ intersects the abscissa. Thus the amount of correction of the bias of the MLE is the distance between the WLE and the MLE, as illustrated with respect to the response pattern in which only one item is correct in Figure 9. It is obvious that the correction makes the estimates of θ regress toward $\theta = b_g = 0.0$ in this example. Note that for the response pattern that consists of 15 correct answers and 15 incorrect answers the correction is nil, and $\tilde{\theta}_v = \hat{\theta}_v$.

[IV. 4] Straight-Forward Methods of Eliminating Bias

Lord (1983) suggested a direct correction of the bias of the MLE for the true test score, which is a monotone transformation of ability θ . When applied to the original ability scale θ this corresponds to $\hat{\theta}_v$ subtracted by $B(\theta;\hat{\theta}_v)$ at $\theta = \hat{\theta}_v$. This correction tends to over-compensate the bias, and a more logical correction may be to identify the value of θ at which the discrepancy from $\hat{\theta}_v$ equals the value of the bias function at that point of θ . This can be done by drawing a line from the $\hat{\theta}_v$ with the angle of 45 degrees from the abscissa until it reaches the curve of the MLE bias function, and then drawing a line vertical to the abscissa. Thus the corrected MLE differs from the original $\hat{\theta}_v$ by the expected amount of bias at that point of θ .

The relationships among the MLE $\hat{\theta}_v$, the two corrected MLE's and Warm's WLE $\tilde{\theta}_v$ are also illustrated in Figure 9. It should be noted that the difference between the two corrected MLE's can be substantially large where the MLE bias function assumes a steep curve.

[IV. 5] Usefulness of Warm's Weight Function as a Prior

These straight-forward corrections of $\hat{\theta}_v$ makes us feel as if Warm's WLE were unnecessary. Note, however, that these two corrected MLE's cannot be obtained either for the *all-correct* response pattern or for the *all-incorrect* response pattern, while Eq. (21) provides WLE's for these extreme response patterns also.

In Bayesian estimation of ability θ , it is customary for researchers in psychology and educational psychology to use the density function representing the ability distribution of some population to which the examinee belongs. Some researchers even believe that, because such a Bayesian estimation of ability uses additional information (i.e., the prior), the resulting ability estimate should be more accurate than the MLE.

This idea contains several serious problems, however. First of all, as Samejima (1969) and

Lord (1986) pointed out, the use of such a prior increases the amount of bias of the ability estimate. Secondly, since such a prior is based on rather trivial factors, such as gender, age, etc., this could lead to serious social and ethical problems. As Lord (1986) stated, the examinee's estimated ability depends not only on his/her test performance but also on the nature of the entire group in which he/she *happens to be included*; if the group as a whole is a low ability group, the examinee's ability estimate may regress downward; if it is a high ability group, his/her estimated ability may regress upward. Thus one may lose a job opportunity if the priors represent gender differences, for example, while he/she may earn the job if they represent socio-economic statuses.

To avoid this, it may be advisable to customize a prior for each individual examinee, by taking the intersection of many different attributes, until finally no one else belongs to the prior than the examinee. If such a prior can be identified, however, there will be no need for testing.

Strengths of the prior used for the WLE are that:

1. the prior is intrinsic in the test and, therefore, nothing beside the examinee's test performance is used in ability estimation, and no unfair discrimination against any individual examinees will arise, and
2. its use will eliminate bias in ability estimation rather than increase it, so it can be used effectively in CAT as well as in paper-and-pencil testing.

V. MODIFIED TEST INFORMATION FUNCTION

In CAT, it has been a widely used practice to adopt a set amount of test information in the stopping rule. That is to say, when the amount of test information of the individually customized subset of items selected from the itempool has reached that criterion amount at which the examinee's individual parameter is currently estimated, no more items will be presented, and testing will be over. This will be legitimate when the individual parameter θ_a of the examinee lies within the interval of θ where the MLE bias function is practically nil, but it requires some modification when it lies outside of this interval.

Samejima proposed two modification formulae of the test information function (see Samejima, 1994b). These modifications are given by

$$\Upsilon(\theta) \;=\; I(\theta)\;[1+\frac{\partial}{\partial\theta}B(\theta;\hat{\theta}_v)]^{-2}$$

and

$$\Xi(\theta) \;=\; I(\theta)\;\{[1+\frac{\partial}{\partial\theta}B(\theta;\hat{\theta}_v)]^2 + I(\theta)\;[B(\theta;\hat{\theta}_v)]^2\}^{-1}\;, \qquad (22)$$

respectively, where $I(\theta)$ is the test information function defined by Eq. (8) and $B(\theta;\hat{\theta}_v)$ is the MLE bias function specified in Eq. (20). This second modified test information function, $\Xi(\theta)$, represents an approximate minimum bound of the mean squared error of the MLE, and the amount of correction is greater for values of θ at which unbiasedness of the MLE is more pronounced. Figure 10 illustrates the square root of the $\Xi(\theta)$ defined by Eq. (22) by a dotted line, in comparison with the square roots of the original test information function $I(\theta)$ and also $\Upsilon(\theta)$ which are drawn by solid and dashed lines, respectively, for the 43 multiple-choice test items of the Iowa Level 11 Vacabulary Subtest, following the logistic model. It will be desirable to use $\Xi(\theta)$ instead of $I(\theta)$ in the stopping rule of CAT, when the MLE is used for the estimate of the examinee's individual parameter.

Insert Figure 10 About Here

VI. DISCUSSION

In this paper, in the effort of improving methods of applying IRT in practical situations, especially in CAT, the content validity of the ability dimension was emphasized, and the idea of core test items was proposed. Devices are proposed to eliminate noise from multiple-choice test items by making use of the nonparametric estimation of operating characteristics effectively in pilot studies, and use of the normal ogive model instead of 3PL was suggested. It was recommended to use several graded response items as those presented at the beginning of CAT in order to avoid the influence of bias and lack of information intrinsic in dichotomous response items. Warm's WLE and its expanded form for general discrete responses were discussed as an effective method of eliminating bias in ability estimation, and the usefulness of Warm's weight function as a prior was discussed. Use of the modified test information function was also suggested for the same purpose.

The author hopes that these methods suggested in the present paper will be tested by researchers in education in actual computerized adaptive testing, and the results will be compared to find out how well each device, or combinations of devices, will work.

References

[1] 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* (chap. 17-20). Reading, MA:Addison Wesley.

[2] Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

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

[4] Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Psychometrika*, *51*, 157-162.

[5] Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading Massachusettes: Addison-Wesley.

[6] Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

[7] Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

[8] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34*, (4, Pt. 2).

[9] Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph No. 18, 37*, (1, Pt. 2).

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

[11] 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.

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

[13] Samejima, F. (1994a). Nonparametric estimation of the plausibility functions of the distractors of vocabulary test items. *Applied Psychological Measurement, 18*, 35-51.

[14] Samejima, F. (1994b). Estimation of the reliability coefficient using the test information function and its two modifications. *Applied Psychological Measurement, 18*, 229-244.

[15] Samejima, F. (1997). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika, 62*, 471-493.

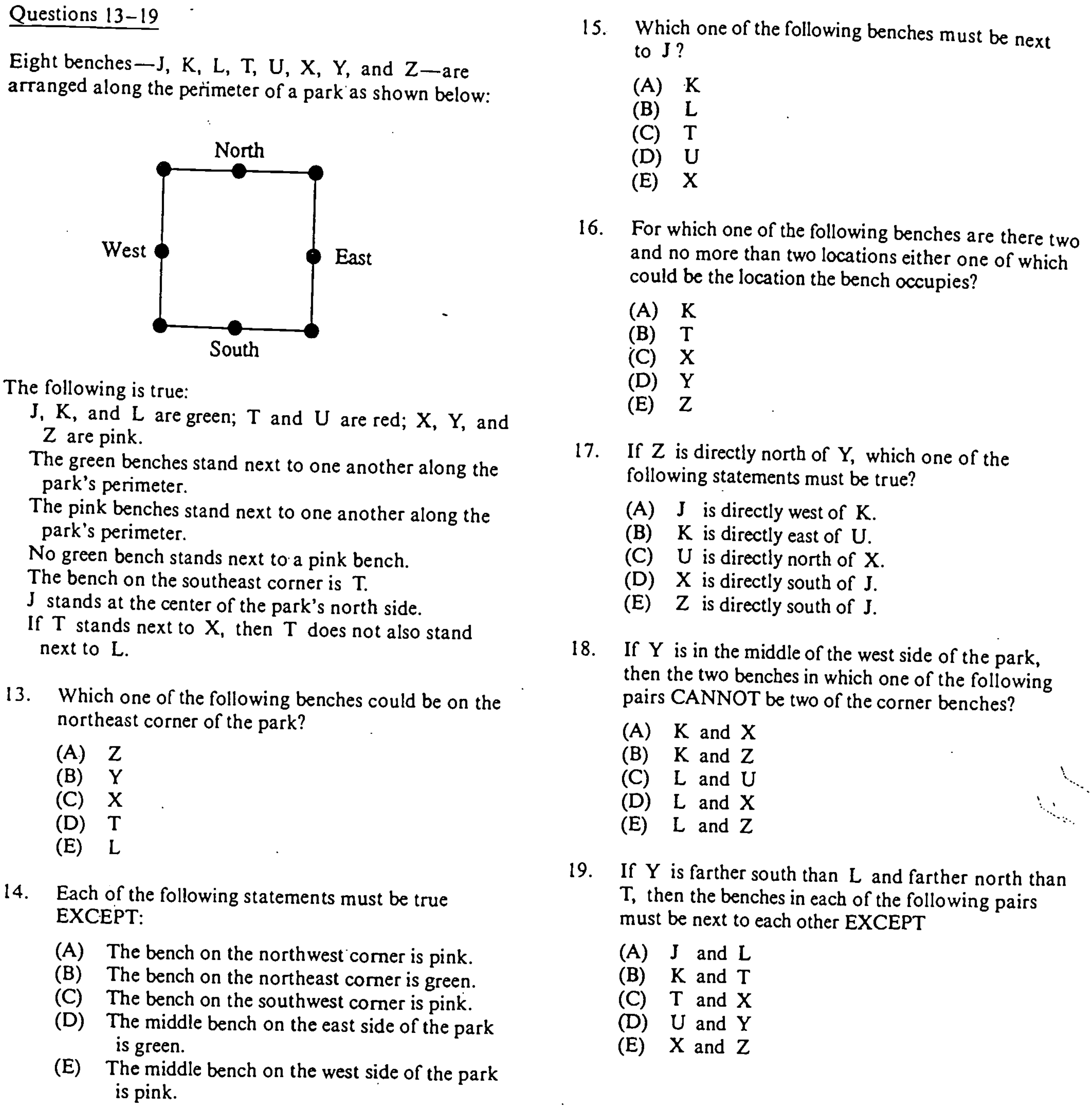
[16] Samejima, F. (1998). Efficient nonparametric approaches for estimating the operating characteristics of discrete item responses. *Psychometrika*, (in press).

[17] Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

[18] Yen, W. M., Burket, G. R. and Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model, *Psychometrika*, *56*, 39-54.

AERA982.TEX

April 5, 1998



Questions 13–19

Eight benches—J, K, L, T, U, X, Y, and Z—are arranged along the perimeter of a park as shown below:

North
West
East
South

The following is true:

J, K, and L are green; T and U are red; X, Y, and Z are pink.

The green benches stand next to one another along the park's perimeter.

The pink benches stand next to one another along the park's perimeter.

No green bench stands next to a pink bench.

The bench on the southeast corner is T.

J stands at the center of the park's north side.

If T stands next to X, then T does not also stand next to L.

13. Which one of the following benches could be on the northeast corner of the park?

(A) Z
(B) Y
(C) X
(D) T
(E) L

14. Each of the following statements must be true EXCEPT:

(A) The bench on the northwest corner is pink.
(B) The bench on the northeast corner is green.
(C) The bench on the southwest corner is pink.
(D) The middle bench on the east side of the park is green.
(E) The middle bench on the west side of the park is pink.

15. Which one of the following benches must be next to J ?

(A) K
(B) L
(C) T
(D) U
(E) X

16. For which one of the following benches are there two and no more than two locations either one of which could be the location the bench occupies?

(A) K
(B) T
(C) X
(D) Y
(E) Z

17. If Z is directly north of Y, which one of the following statements must be true?

(A) J is directly west of K.
(B) K is directly east of U.
(C) U is directly north of X.
(D) X is directly south of J.
(E) Z is directly south of J.

18. If Y is in the middle of the west side of the park, then the two benches in which one of the following pairs CANNOT be two of the corner benches?

(A) K and X
(B) K and Z
(C) L and U
(D) L and X
(E) L and Z

19. If Y is farther south than L and farther north than T, then the benches in each of the following pairs must be next to each other EXCEPT

(A) J and L
(B) K and T
(C) T and X
(D) U and Y
(E) X and Z

FIGURE 1

Taken from LSAT, the Official Prep Test III, 1991, Vol. 2, Page 77: An Example Question.

(a) J, K, and L are green; T and U are red; X, Y, and Z are pink.

(b) The green benches stand next to one another along the park's perimeter.

(c) The pink benches stand next to one another along the park's perimeter.

(d) No green bench stands next to a pink bench.

(e) The bench on the southeast corner is T.

(f) J stands at the center of the park's north side.

(g) If T stands next to X, then T does not also stand next to L.

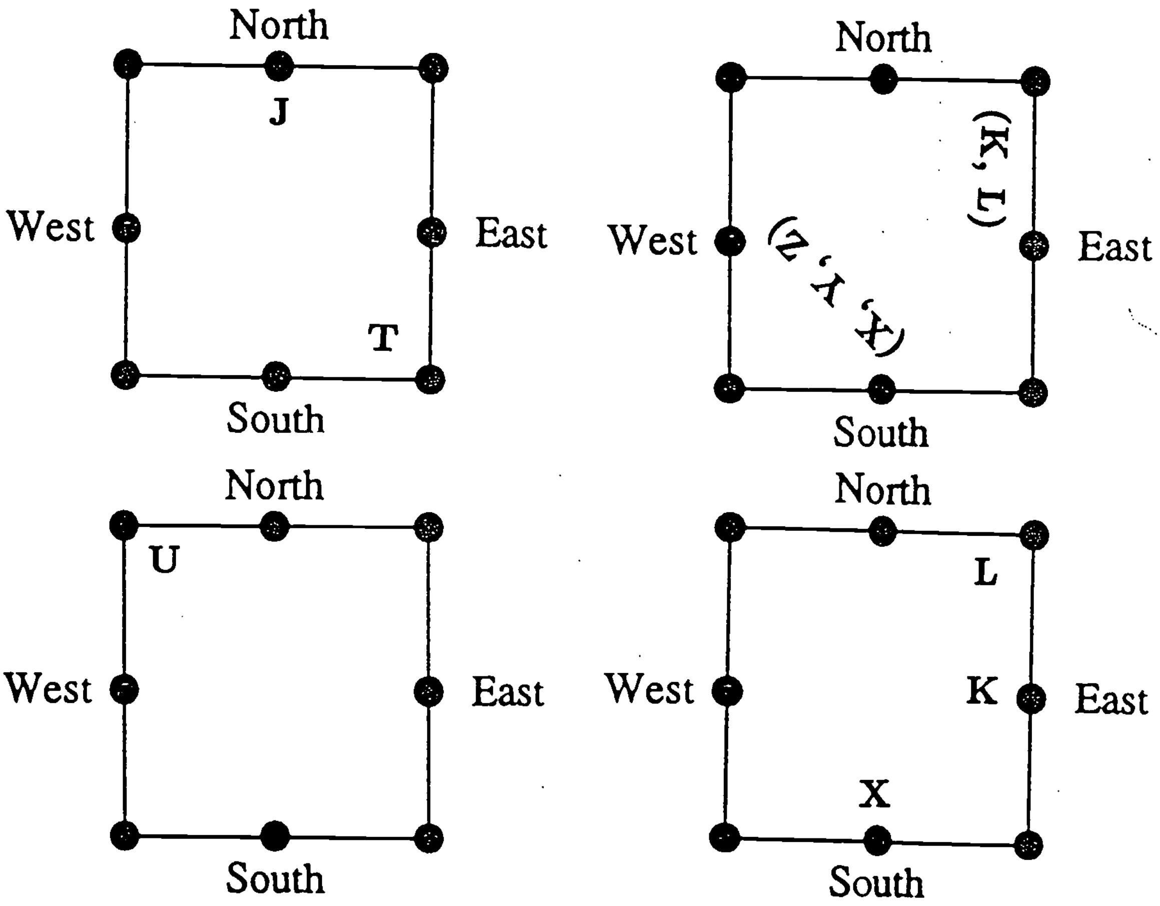


FIGURE 2

Modified LSAT Example to a Graded Response Item.

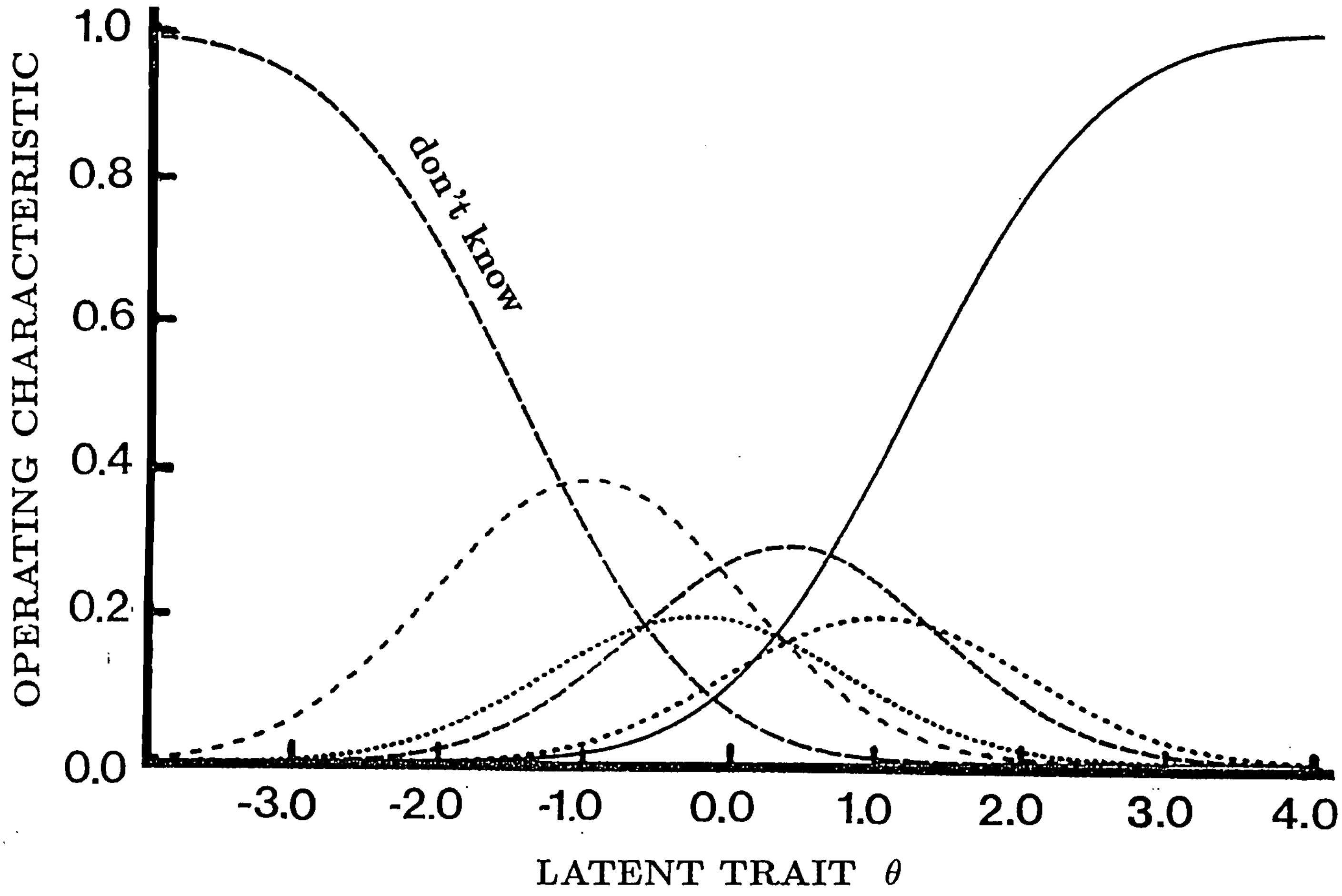


FIGURE 3

Normal Ogive Model; $a_g = 1.0$, $b_{x_g} = -1.50, -0.50, 0.00, 0.75, 1.25$

(Taken from ONR/RR-79-4: page 18)

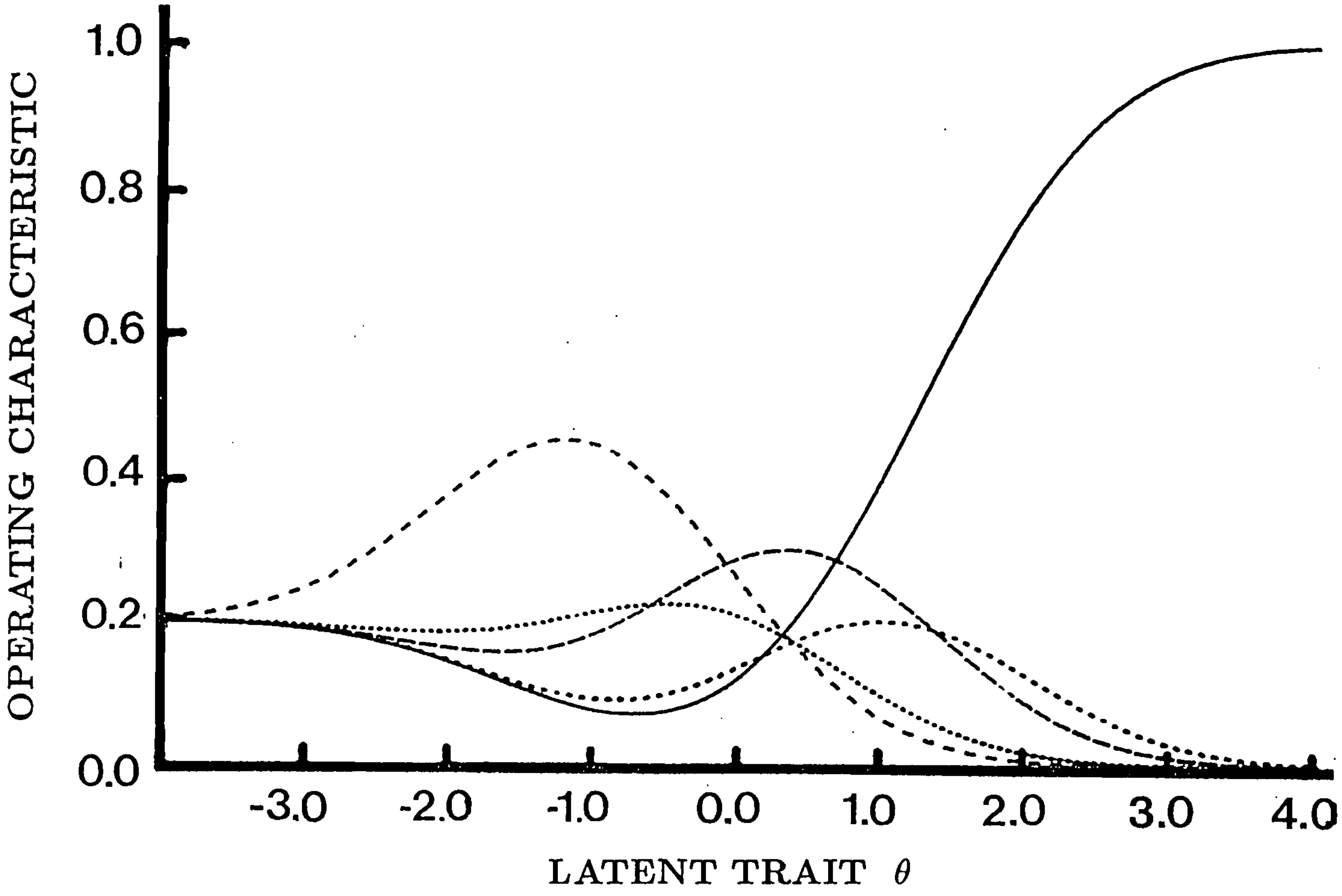


FIGURE 4

Model A; $a_g = 1.0$, $b_{x_g} = -1.50, -0.50, 0.00, 0.75, 1.25$

(Taken from ONR/RR-79-4: page 20)

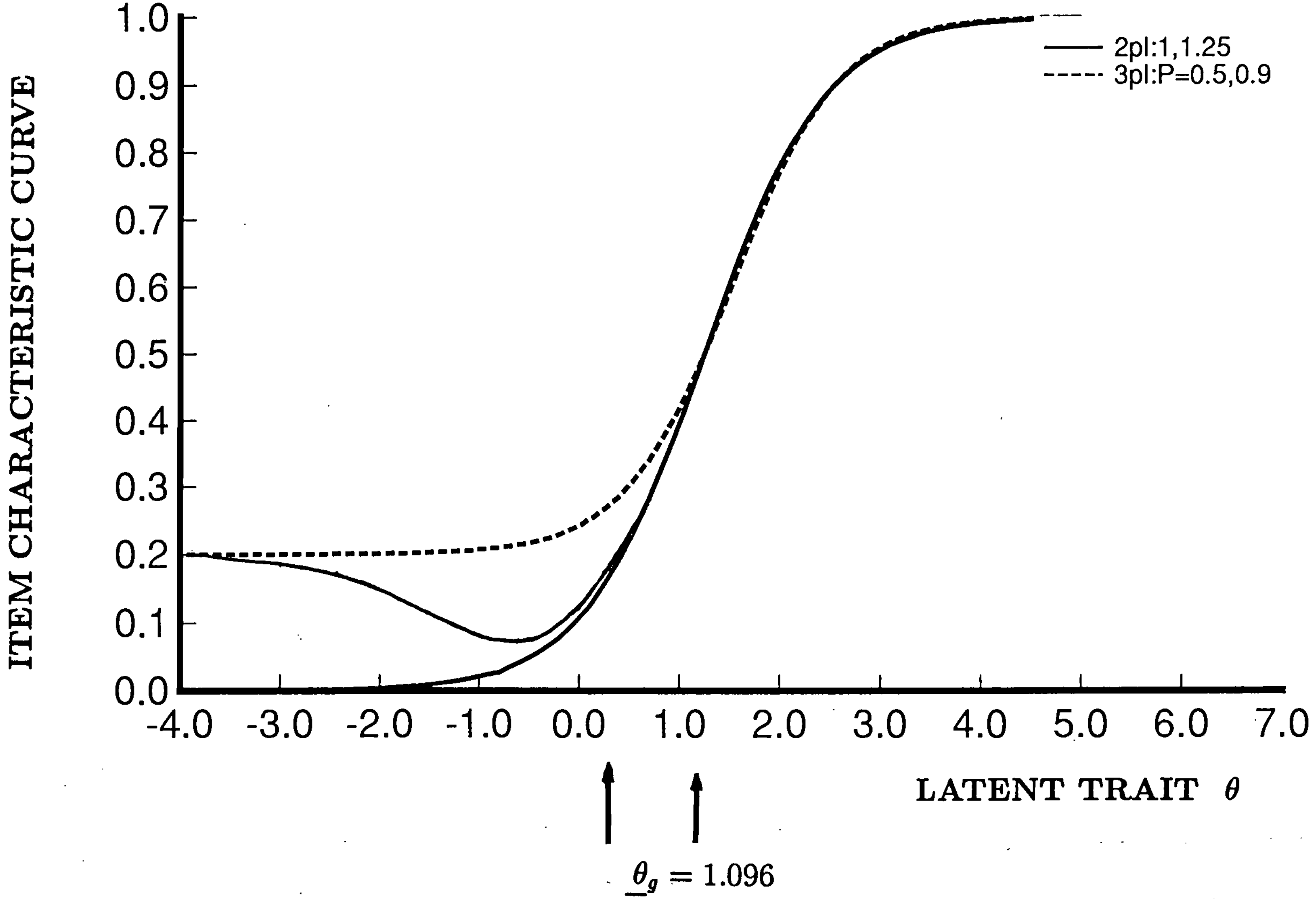


FIGURE 5

ICC Affected by the Plausibility Functions of Distractors in Comparison with the 2-Parameter and 3-Parameter Logistic ICC's.

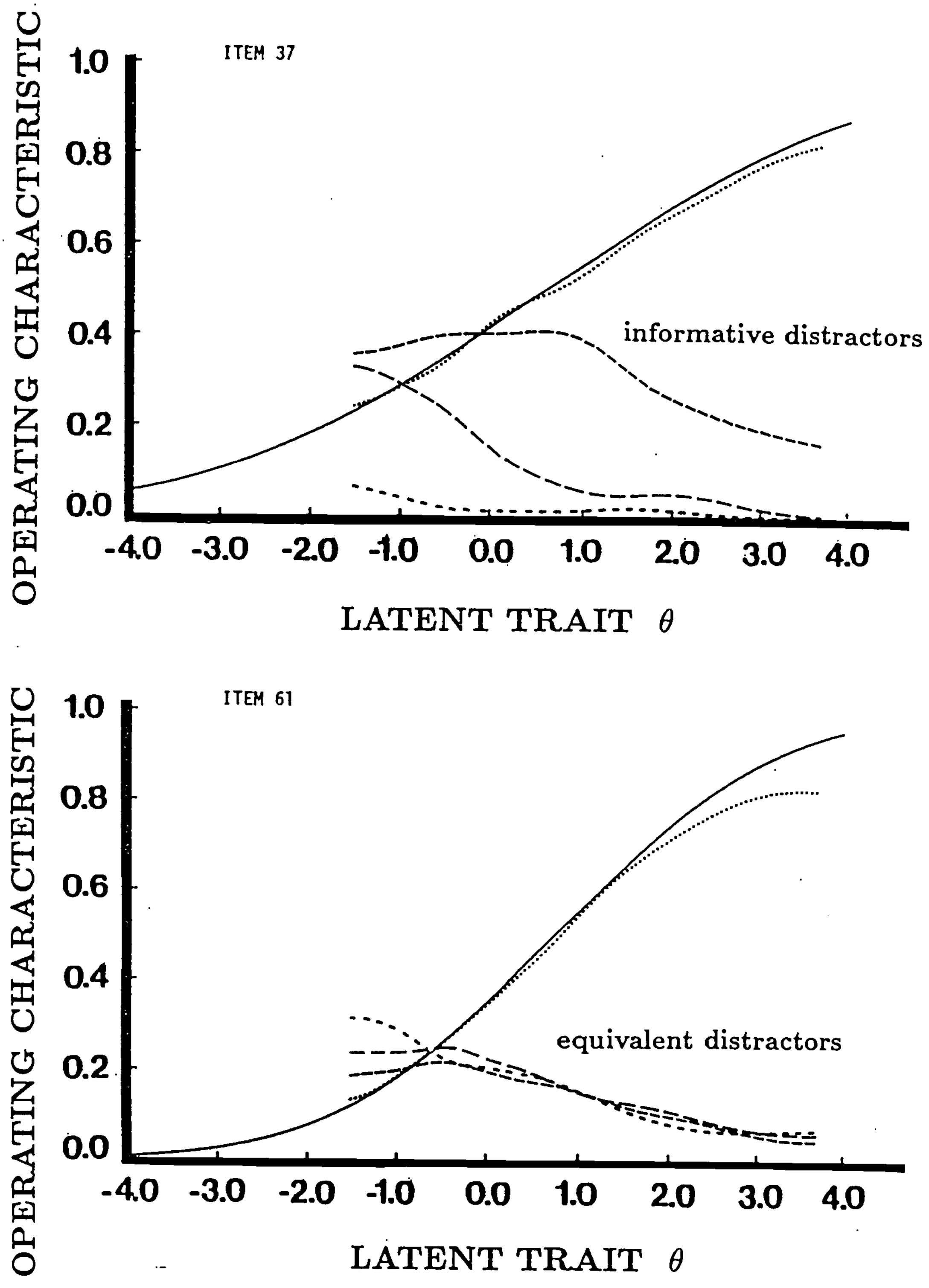


FIGURE 6

Examples of Informative Distractors and Equivalent Distractors

(Taken from ONR/RR-84-1: pages 43 and 55.)

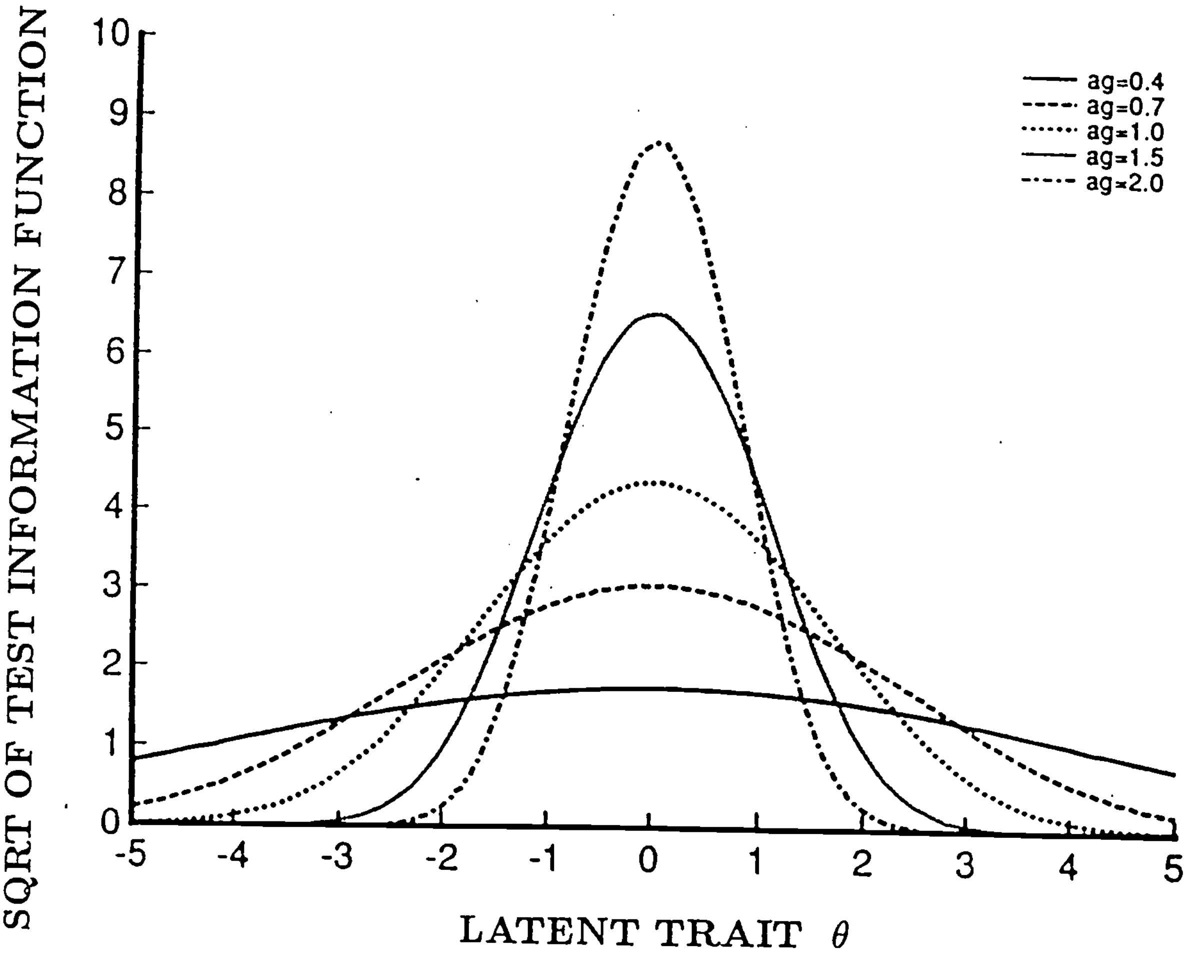


FIGURE 7

Square Root of the Test Information Function of Each of the Five Hypothetical Tests of 30 Equivalent Items: the Normal Ogive Model.

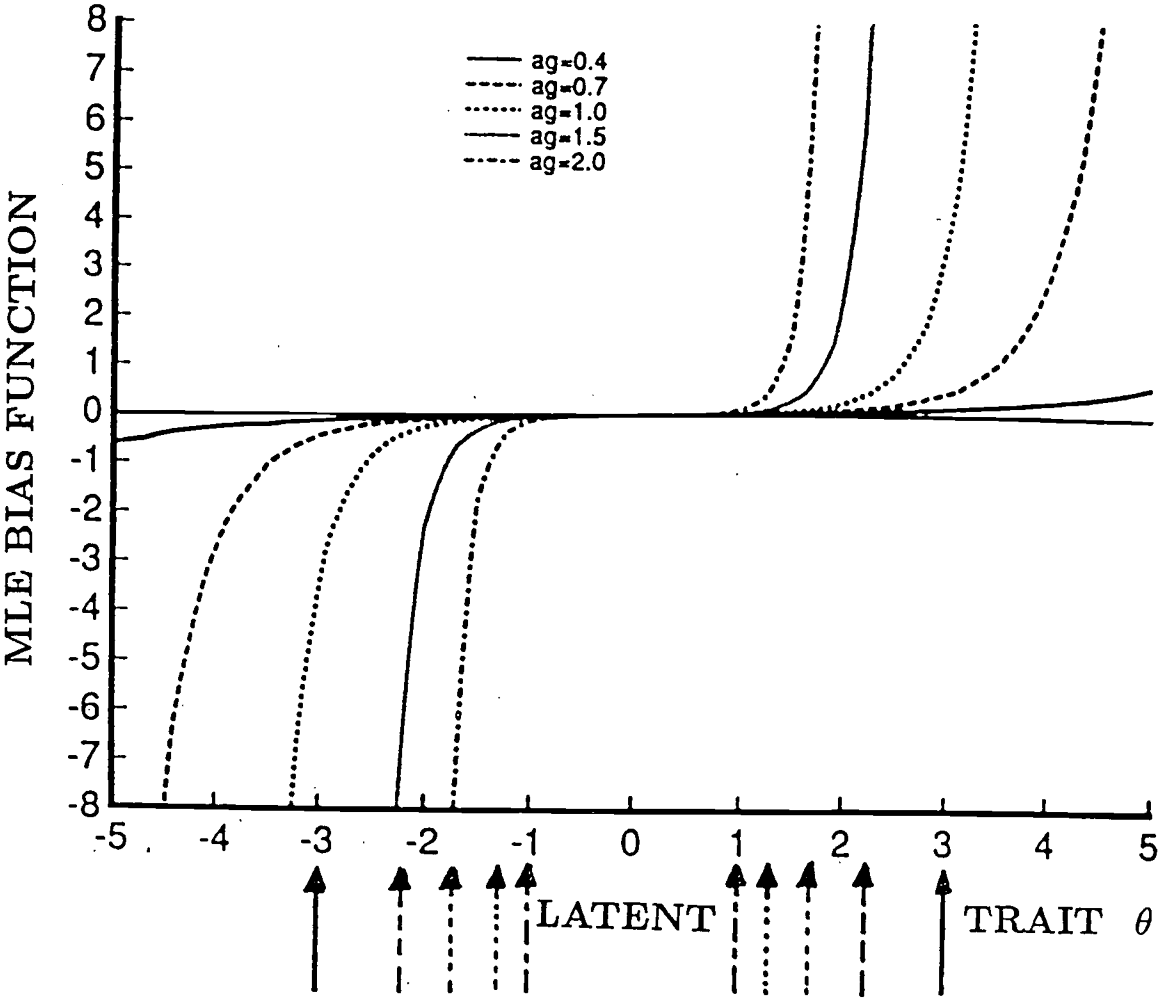


FIGURE 8

MLE Bias Function of Each of the Five Hypothetical Tests of 30 Equivalent Items: the Normal Ogive Model.

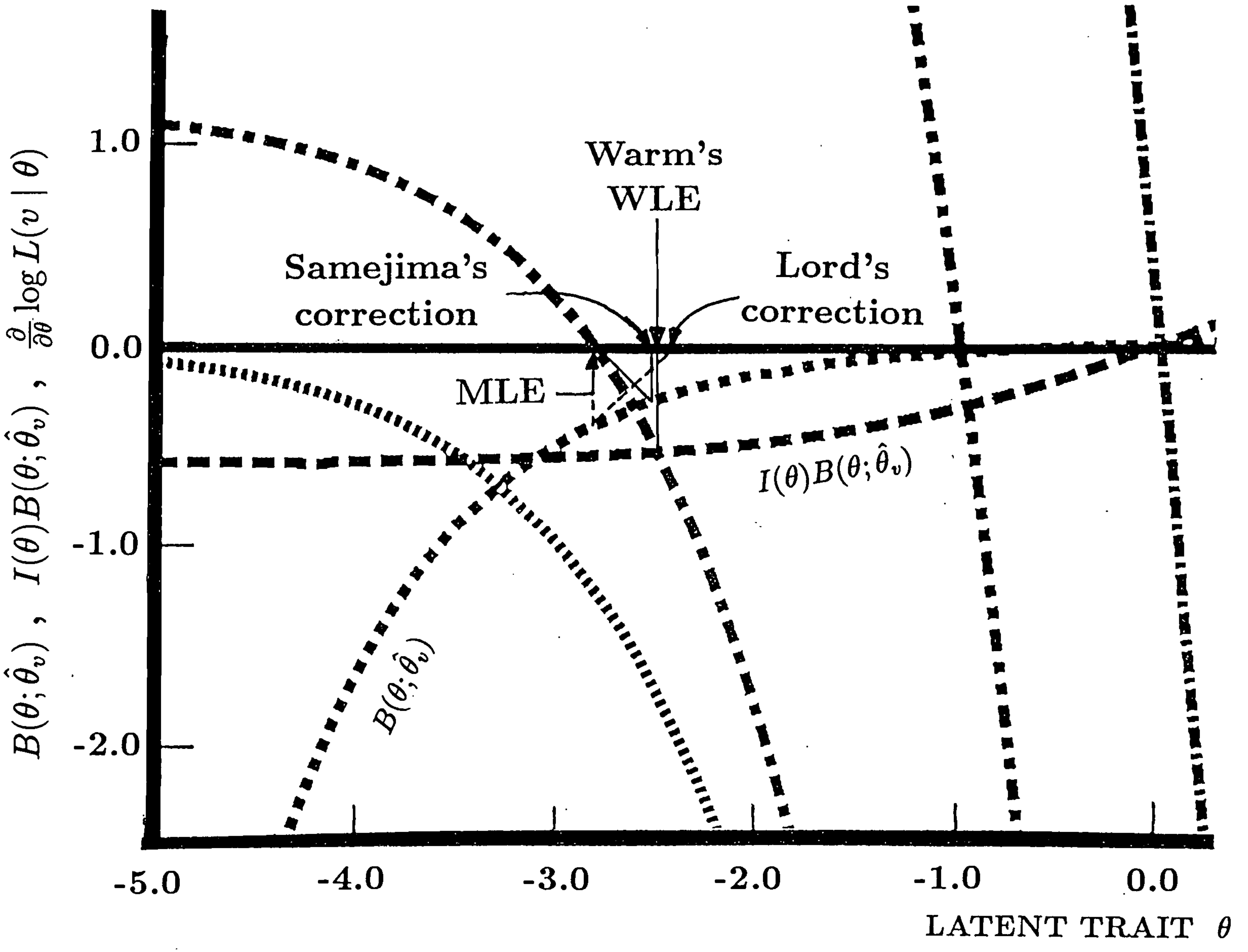
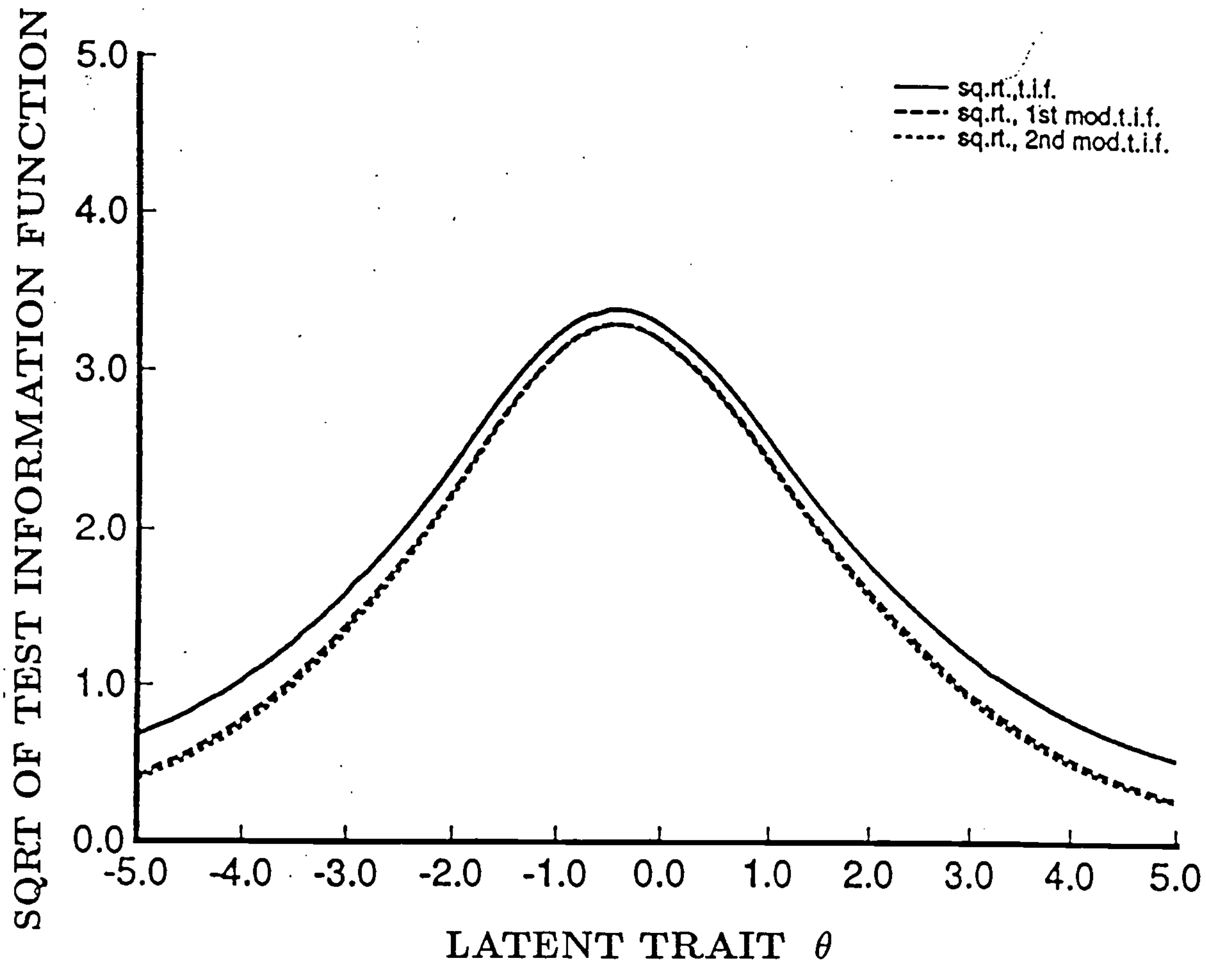


FIGURE 9

Relationships among the MLE $\hat{\theta}_v$, the Two Corrected MLE's by Lord and Samejima, Respectively, and Warm's WLE $\tilde{\theta}_v$.



0.750 0.50 1.50 6.00 6.00
8007HOL_DAT, IN9007, plotted by NANCY DOMM

FIGURE 10

Square Roots of the Test Information Function (Solid) and Its Two Modifications (Dashed & Dotted) of the Iowa Level 11 Vocabulary Subtest: the Logistic Model.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031262

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: SOME CONSIDERATIONS FOR ELIMINATING BIASES IN ABILITY ESTIMATION IN COMPUTERIZED ADAPTIVE TESTING	
Author(s): DR. FUMIKO SAMEJIMA	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce end disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Fumiko Samejima	Printed Name/Position/Title: DR. FUMIKO SAMEJIMA, PROFESSOR	
Organization/Address: THE UNIVERSITY OF TENNESSEE 405 AUSTIN PEAY BUILDING, DEPT OF PSYCHOLOGY KNOXVILLE, TN 37996 - 0900	Telephone: (423) 974 - 3008 E-Mail Address: Samejima@psych1.psych.utk.edu	FAX: (423) 974 - 3330 Date: 4/1/98

(over)

ERIC®

Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

March 20, 1998

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae.net.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (424)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1998/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an AERA chair or discussant, please save this form for future use.

CUA

The Catholic University of America