DOCUMENT RESUME

ED 325 515 TM 015 764

AUTHOR Samejima, Fumiko
TITLE Validity Study in Multidimensional Latent Space and Efficient Computerized Adaptive Testing. Final Report.
INSTITUTION Tennessee Univ., Knoxville. Dept. of Psychology.
SPONS AGENCY Office of Naval Research, Arlington, VA. Cognitive and Neural Sciences Div.
PUB DATE 24 Sep 90
CONTRACT ONR-N00014-87-K-0320
NOTE 90p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Distractors (Tests); Equations (Mathematics); Estimation (Mathematics); Evaluation Research; *Federal Programs; Graphs; *Item Response Theory; Mathematical Models; Nonparametric Statistics; *Research Projects; *Test Validity
IDENTIFIERS Final Reports; Information Function (Tests); *Multidimensional Latent Space

ABSTRACT

This paper is the final report of a multi-year project sponsored by the Office of Naval Research (ONR) in 1987 through 1990. The main objectives of the research summarized were to: investigate the non-parametric approach to the estimation of the operating characteristics of discrete item responses; revise and strengthen the package computer programs and implement them in the Unix Operating System; investigate computerized adaptive testing procedure and use it in the SUN microcomputer system networked with personal computers; investigate multidimensional latent trait theory; and study item validity and test validity using the multidimensional latent space. Products published or presented during the research period included: five research reports through the ONR; a special contribution paper, "Comprehensive Latent Trait Theory"; 13 papers presented at conferences; and other seminars and research collaborations. This report reviews: (1) backgrounds and basic concepts used throughout the research; (2) two formulae for modification of the test information function; (3) the reliability coefficient and standard error of measurement in classical test theory in the context of latent trait models; (4) validity measures in the context of latent trait models; (5) the non-parametric approach to estimation of the operating characteristics of discrete item responses; (6) content-based observation of informative distractors and the efficiency of ability estimation; and (7) efficient computerized adaptive testing. Thirty graphs and five data tables are included. A 167-item distribution list is appended. (SLD)

VALIDITY STUDY IN MULTIDIMENSIONAL LATENT SPACE AND EFFICIENT COMPUTERIZED ADAPTIVE TESTING

FUMIKO SAMEJIMA

UNIVERSITY OF TENNESSEE

KNOXVILLE, TENN. 37996-0900

SEPTEMBER, 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

R01-1069-11-0[illegible]-91

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1a REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; Distribution unlimited
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Fumiko Samejima, Ph.D. Psychology Department 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION Cognitive Science 1142 CS
6c ADDRESS (City, State, and ZIP Code) 310B Austin Peay Building The University of Tennessee Knoxville, TN 37996-0900	7b ADDRESS (City, State, and ZIP Code) Office of Naval Research 800 N. Quincy Street Arlington, VA 22217
8a NAME OF FUNDING / SPONSORING ORGANIZATION Cognitive Science Research Program 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-87-K-0320
8c ADDRESS (City, State, and ZIP Code) Office of Naval Research 800 N. Quincy Street Arlington, VA 22217	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61153N PROJECT NO: RR-042-04 TASK NO: 042-04-01 WORK UNIT ACCESSION NO: 4421-549
11 TITLE (Include Security Classification) Validity study in multidimensional latent space and efficient computerized adaptive testing	
12 PERSONAL AUTHOR(S) Fumiko Samejima, Ph.D.	
13a TYPE OF REPORT final report	13b TIME COVERED FROM 1987 TO 1990 14 DATE OF REPORT (Year, Month, Day) September 24, 1990 15 PAGE COUNT 88
16 SUPPLEMENTARY NOTATION	

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Latent Trait Models, Mental Test Theory, Multiple-Choice Test, Computerized Adaptive Testing, Test Reliability, Test Validity, Test Information Function, Nonparametric Estimation

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This is a summary of the research conducted in the past three years and seven months, 1987-90, under the title, "Validity Study in Multidimensional Latent Space and Efficient Computerized Adaptive Testing."

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION
22a NAME OF RESPONSIBLE INDIVIDUAL Dr. Charles E. Davis	22b TELEPHONE (Include Area Code) 202-696-4046 22c OFFICE SYMBOL ONR-1142-CS

DD Form 1473, JUN 86 Previous editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

S/N 0102-LF-014-6603

PREFACE

Three and a half years have passed since I started this research on March 1, 1987. During this period, so many things were designed and accomplished, and as the principal investigator I find it extremely difficult to include and systematize all the important findings and implications within a single final report. It is my regret that many of them have to be left out, but I did my best within a limited amount of time with the hope that this final report will help the reader to grasp the outline of the whole accomplishment.

There were five main objectives in the original research proposal, and they can be summarized as follows.

[1] Further investigate the nonparametric approach to the estimation of the operating characteristics of discrete item responses.

[2] Revise and strengthen the package computer programs and eventually implement them in the Unix Operating System.

[3] Investigate an ideal computerized adaptive testing procedure and eventually materialize it in the SUN microcomputer system networked with IBM personal computers.

[4] Investigate multidimensional latent trait theory.

[5] Pursue item validity and test validity using the multidimensional latent space.

Out of these objectives, Objectives [1] and [5], together with Objectives [2] and [3], were most intensively pursued. The highest productivity belongs to this part of the research, which provides us with valuable future perspectives of research.

During the research period there were many people who helped me as assistants, secretaries, etc., as I acknowledged in each research report. Also people of the Office of Naval Research, especially Dr. Charles E. Davis, and those of the ONR Atlanta Office, including Mr. Thomas Bryant, have been of great help in conducting the research. I would like to express my gratitude to all of them.

Thanks are also due to my assistants, Nancy H. Domm and Raed A. Hijer, who helped me in preparing this final report. Appreciation is also extended to my former assistants, Christine A. Golik and Philip S. Livingston, who still helped me occasionally during the research period.

September 20, 1990

Author

TABLE OF CONTENTS

	Page
I Introduction	1
I.1 Research Reports	1
I.2 Special Contribution Paper	1
I.3 Paper Presentations at Conferences	1
I.4 Other Events	2
II Backgrounds and Basic Concepts Used throughout the Research	3
II.1 General Concepts in Latent Trait Models	3
II.2 Critical Observations of the Reliability, Standard Error of Measurement and Validity of a Test	4
II.3 Nonparametric Approach to the Estimation of the Operating Characteristics of Discrete Item Responses	4
II.4 Possible Non-Monotonicities of the Operating Characteristics	7
III Proposal of Two Modification Formulae of the Test Information Function	9
III.1 Minimum Variance Bound	10
III.2 First Modified Test Information Function	11
III.3 Minimum Bound of the Mean Squared Error	12
III.4 Second Modified Test Information Function	13
III.5 Examples	13
III.6 Minimum Bounds of Variance and Mean Squared Error for the Transformed Latent Variable	22
III.7 Modified Test Information Functions Based upon the Transformed Latent Variable	24
III.8 Discussion and Conclusions	25

IV Reliability Coefficient and Standard Error of Measurement in Classical Mental Test Theory Predicted in the Context of Latent Trait Models 26

IV.1 General Case 26

IV.2 Reliability Coefficient of a Test in the Sense of Classical Mental Test Theory When the Maximum Likelihood Estimator of θ Is Used 27

IV.3 Standard Error of Measurement of a Test in the Sense of Classical Mental Test Theory When the Maximum Likelihood Estimator of θ Is Used 28

IV.4 Examples 29

IV.5 Discussion and Conclusions 33

V Validity Measures in the Context of Latent Trait Models 34

V.1 Performance Function: Regression of the External Criterion Variable on the Latent Variable 34

V.2 When $\varsigma(\theta)$ Is Strictly Increasing in θ : Simplest Case 35

V.3 Test Validity Measures Obtained from More Accurate Minimum Variance Bounds 41

V.4 Multidimensional Latent Space 42

V.5 Discussion and Conclusions 44

VI Further Investigation of the Nonparametric Approach to the Estimation of the Operating Characteristics of Discrete Item Responses 44

VI.1 Simple Sum Procedure of the Conditional P.D.F. Approach Combined with the Normal Approach Method 45

VI.2 Differential Weight Procedure 45

VI.3 Examples 47

VI.4 Sensitivities to Irregularities of Weight Functions 47

VI.5 Discussion and Conclusions 49

VII Content-Based Observation of Informative Distractors and Efficiency of Ability Estimation 52

VII.1 Non-Monotonicity of the Conditional Probability of the Positive Response, Given Latent Variable 52

VII.2 Effect of Noise in the Three-Parameter Logistic Model and the Meanings of the Difficulty and Discrimination Parameters 59

VII.3 Informative Distractors of the Multiple-Choice Test Item 62

VII.4 Merits of the Nonparametric Approach for the Identification of Informative Distractors and for the Estimation of the Operating Characteristics of an Item 64

VII.5 Efficiency in Ability Estimation and Strategies of Writing Test Items 64

VII.6 Discussion and Conclusions 69

VIII Efficient Computerized Adaptive Testing 70

VIII.1 Validity Measures Tailoring a Sequential Subset of Items for an Individual 70

VIII.2 Use of the Modifications of the Test Information Function in Stopping Rules 70

VIII.3 Use of Test Validity Measures in Stopping Rules 71

VIII.4 Prediction of the Reliability Coefficient for a Specific Population of Examinees in Computerised Adaptive Testing 71

VIII.5 Differential Weight Procedure for Item Analysis and for On-Line Item Calibration 72

VIII.6 Use of Informative Distractors 73

VIII.7 Discussion and Conclusions 73

IX Other Findings in the Present Research 74

I Introduction

This is the final report of the multi-year research project entitled *Validity Study in Multidimensional Latent Space and Efficient Computerized Adaptive Testing*, which was sponsored by the Office of Naval Research in 1987 through 1990 (N00014-87-K-0320). The accomplishments include those which have already been published as ONR research reports as well as those still in progress, which will be published in later years as part of more comprehensive research results.

The rest of this chapter will describe papers published or presented during the research period, and related events. The contents of the research accomplishments will be summarized and systematized, and will be described in the succeeding chapters.

[I.1] Research Reports

The following are the ONR research reports that have been published in the present research project.

(1) Modifications of the Test Information Function. *Office of Naval Research Report 90-1*, 1990.

(2) Predictions of Reliability Coefficients and Standard Errors of Measurement Using the Test Information Function and its Modifications. *Office of Naval Research Report 90-2*, 1990.

(3) Validity Measures in the Context of Latent Trait Models. *Office of Naval Research Report 90-3*, 1990.

(4) Differential Weight Procedure of the Conditional P.D.F. Approach for Estimating the Operating Characteristics of Discrete Item Responses. *Office of Naval Research Report 90-4*, 1990.

(5) Content-Based Observation of Informative Distractors and Efficiency of Ability Estimation. *Office of Naval Research Report 90-5*, 1990.

[I.2] Special Contribution Paper

During this period, with the request of Dr. Chikio Hayashi, president of the Behaviormetric Society, a special contribution paper entitled *Comprehensive Latent Trait Theory* was written and published in *Behaviormetrika*, Vol. 24, 1988. The paper is based upon the invited address, a one hour special lecture overviewing latent trait models, which was given at the 1987 Annual Meeting of the Behaviormetric Society in 1987 at Kyushu University, Fukuoka, Japan, under the title, *Overview of Latent Trait Models*. There were more than two hundred researchers in the audience, and the summary of the paper is given as Appendix B of the author's ONR Final Report: Advancement of Latent Trait Theory, which was published in 1988.

[I.3] Paper Presentations at Conferences

There are thirteen papers presented at conferences during this research period, *excluding* those in 1987 which have been reported in " Final Report: Advancement of Latent Trait Theory." They include ONR contractors' meetings, and are listed below.

(1) *A Robust Method of On-Line Calibration.* American Educational Research Association Meeting, New Orleans, 1988. U. S. A.

(2) *Some Modifications of the On-Line Item Calibration Methods.* ONR Conference on Model-Based Measurement, Iowa City, 1988. U. S. A.

(3) *Information Functions of the General Model Developed for Differential Strategies and Possibilities for Applying Half-Discrete, Half-Continuous Models for Projective Techniques.* ONR Conference on Model-Based Measurement, Iowa City, 1988. U. S. A.

(4) *Some Refinement in the Estimation of the Operating Characteristics of Discrete Item Responses without Assuming any Mathematical Form.* Psychometric Society Meeting, Los Angeles, 1988. U. S. A.

(5) *Prospect of Analyzing Rorschach Data by Sophisticated Psychometric Methods.* Symposium: The Burstein-Loucks Rorschach Scoring System: Clinical and Psychometric Developments. American Psychological Association Annual Meeting, Atlanta, 1988. U. S. A.

(6) *Latent Trait Approach to Rorschach Diagnosis Based upon the Burstein-Loucks Scoring System.* American Educational Research Association Annual Meeting, San Francisco, 1989. U. S. A. (round-table session)

(7) *Some Considerations on Validity Measures in Latent Trait Theory.* ONR Conference on Model-Based Measurement, Norman, OK, 1989. U. S. A.

(8) *Differential Weight Procedure of the Conditional P.D.F. Approach in the Estimation of Operating Characteristics of Discrete Item Responses.* ONR Conference on Model-Based Measurement, Norman, OK, 1989. U. S. A.

(9) *Some Reliability and Validity Measures in the Context of Latent Trait Models.* Psychometric Society Annual Meeting, Los Angeles, 1989. U. S. A.

(10) *Prospect of Applying Latent Trait Models and Methodologies Accomodating Both Psychological and Neurological Factors.* American Educational Research Association Annual Meeting, Boston, 1990. U. S. A.

(11) *Reliability/Validity Indices in the Context of Latent Trait Models.* American Educational Research Association Annual Meeting, Boston, 1990. U. S. A.

(12) *Further Considerations for the Differential Weight Procedure of Estimating the Operating Characteristics of Discrete Item Responses.* ONR Conference on Model-Based Measurement, Portland, OR, 1990. U. S. A.

(13) *Modified Test Information Functions, Their Usefulnesses and Prediction of the Test Reliability Coefficient Tailored for a Specific Ability Distribution.* ONR Conference on Model-Based Measurement, Portland, OR, 1990. U. S. A.

[I.4] Other Events

The principal investigator gave a seminar entitled *Comprehensive Latent Trait Models* in September, 1989, at the National Center for University Entrance Examination, Tokyo, Japan, invited by Dr. Shuichi Iwatsubo of the Center and Dr. Kazuo Shigematsu of the Tokyo Engineering University.

She also made research collaborations with Professor Sukeyori Shiba of the University of Tokyo, and with Dr. Takahiro Sato of the C & C Information Technology Research Laboratories of Nippon Electric Company, Japan.

II Backgrounds and Basic Concepts Used throughout the Research

In this chapter, the backgrounds and the basic concepts upon which the present research has been conducted are introduced. The reader is directed to the auther's two previous ONR final reports (Samejima, 1981b, 1988) and other ONR research reports, if he/she wants to know these concepts and developments in more detail.

[II.1] General Concepts in Latent Trait Models

Let θ be *ability*, or *latent trait*, which assumes any real number. Let $g\ (= 1, 2, \cdots, n)$ denote an *item*, k_g be any 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 θ, i.e.,

$$P_{k_g}(\theta) = prob.[k_g \mid \theta] \quad . \tag{2.1}$$

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

$$I_{k_g}(\theta) = -\frac{\partial^2}{\partial\theta^2}\log P_{k_g}(\theta) = [\frac{\partial}{\partial\theta}P_{k_g}(\theta)\ \{P_{k_g}(\theta)\}^{-1}]^2 - \frac{\partial^2}{\partial\theta^2}P_{k_g}(\theta)\ [P_{k_g}(\theta)]^{-1} \quad , \tag{2.2}$$

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

$$I_g(\theta) = E[I_{k_g}(\theta) \mid \theta] = \sum_{k_g} I_{k_g}(\theta)P_{k_g}(\theta) = \sum_{k_g} [\frac{\partial}{\partial\theta}P_{k_g}(\theta)]^2[P_{k_g}(\theta)]^{-1} \quad . \tag{2.3}$$

In the special case where the item g is scored dichotomously, this item information function is simplified to become

$$I_g(\theta) = [\frac{\partial}{\partial\theta}\ P_g(\theta)]^2\ [\{P_g(\theta)\}\{1 - P_g(\theta)\}]^{-1} \quad , \tag{2.4}$$

where $P_g(\theta)$ denotes the operating characteristic of the correct answer to item g.

Let V be a response pattern such that

$$V = \{\ k_g\ \}' \qquad g = 1, 2, ..., n \quad . \tag{2.5}$$

The operating characteristic, $P_V(\theta)$, of the response patten V is defined as the conditional probability of V, given θ. Throughout this report the principle of *local independence* is assumed to be valid, so that within any group of examinees all characterised by the same value of the latent variable θ the distributions of the item response categories are all independent of each other. Thus the operating characteristic of a given response pattern is a product of the operating characteristics of the item response categories contained in that response pattern, so that we can write

$$P_V(\theta) = \prod_{k_g \in V} P_{k_g}(\theta) \quad . \tag{2.6}$$

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

$$I_V(\theta) = -\frac{\partial^2}{\partial\theta^2}\log P_V(\theta) = \sum_{k_g \in V} I_{k_g}(\theta) \quad , \tag{2.7}$$

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

$$I(\theta) = E[I_V(\theta) \mid \theta] = \sum_V I_V(\theta) P_V(\theta) = \sum_{g=1}^{n} I_g(\theta) \quad . \tag{2.8}$$

[II.2] Critical Observations of the Reliability, Standard Error of Measurement and Validity of a Test

The reliability coefficient and the standard error of measurment in classical mental test theory are two concepts that have widely been accepted and used by psychologists and test users in the past decades. The author has pointed out repeatedly, however, that these measures are actually the attributes of a specified group of examinees as well as of a given test. In addition, even if we take this fact into account, 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. In contrast to this, in latent trait models, the item and test information functions, which are defined by (2.3) and (2.8), respectively, provide us with abundant information about the *local* accuracy of estimation, a concept which is totally missing in classical mental test theory. These functions are population-free, i.e., they do not depend upon any specific group of examinees as the reliability coefficient and the standard error of measurment do.

Unlike the progressive dissolution of test reliability, test validity is one concept that has rather been neglected in the context of latent trait models. Several types of validity have been identified and discussed in classical mental test theory, which include content validity, construct validity, and criterion-oriented validity. Perhaps we can say that, in modern mental test theory, both content validity and construct validity are well accomodated, although they are not explicitly stated. If each item is based upon cognitive processes that are directly related to the ability to be measured, then the content of the operationally defined latent variable behind the examinees' performances will be validated. Also construct validity can be identified, with all the mathematically sophisticated structures and functions which characterise latent trait models and which classical mental test theory does not provide. With respect to the criterion-oriented validity, however, so far latent trait models have not offered so much as they did to the test reliability and to the standard error of measurement.

In classical mental test theory, the validity coefficient is again a single number, i.e., the product-moment correlation coefficient between the test score and the criterion variable. Since the correlation coefficient is largely affected by the heterogeneity of the group of examinees, i.e., for a fixed test the coefficient tends to be higher when individual differences among the examinees in the group are greater, and vice versa (cf. Samejima, 1977b), we must keep in mind that *so-called* test validity represents the degree of heterogeneity in ability among the examinees tested, as well as the quality of the test itself.

[II.3] Nonparametric Approach to the Estimation of the Operating Characteristics of Discrete Item Responses

As early as in 1977 the author proposed Normal Approximation Method (Samejima, 1977b) which can be used for item calibration both in computerised adaptive testing and in paper-and-pencil testing. She also discussed the effective use of information functions in adaptive testing (Samejima, 1977a). Since then, with the support by the Office of Naval Research, she has developed several approaches and methods for the same purpose (cf. Samejima, 1977c, 1978a, 1978b, 1978c, 1978d, 1978e, 1978f, 1980a,

1980b, 1981a, 1981b, 1988; Samejima and Changas, 1981). For convenience, they can be categorised as follows.

Approaches	Methods
(1) Bivariate P.D.F. Approach	(1) Pearson System Method
(2) Histogram Ratio Approach	(2) Two-Parameter Beta Method
(3) Curve Fitting Approach	(3) Normal Approach Method
(4) Conditional P.D.F. Approach	(4) Lognormal Approach Method
(4.1) Simple Sum Procedure	
(4.2) Weighted Sum Procedure	
(4.3) Proportioned Sum Procedure	

Here by an approach we mean a general procedure in approaching the operating characteristics of a discrete item response, and by a method we mean a specific method in approximating the conditional density of ability, given its maximum likelihood estimate. Thus a combination of an approach and a method provides us with a specific procedure for estimating the operating characteristic of a discrete item response.

These approaches and methods are characterised by two features, i.e.,

(1) estimation is made without assuming any mathematical forms for the operating characteristics of discrete item responses, and

(2) estimation is efficient enough to base itself upon a relatively small set of data of, say, several hundred to a few thousand examinees.

The backgrounds common to the Bivariate and Conditional Approaches and the differences among different methods can be described as follows. For the sake of simplicity in handling mathematics, the tentative transformation of θ to τ is made by

$$\tau = C_1^{-1} \int_{-\infty}^{\theta} [I(t)]^{1/2} dt + C_0 \quad , \tag{2.9}$$

where C_0 is an arbitrary constant for adjusting the origin of τ, and C_1 is an arbitrary constant which equals the square root of the test information functions, $I^*(\tau)$, of τ, so that we can write

$$C_1 = [I^*(\tau)]^{1/2} \tag{2.10}$$

for all τ. This transformation will be simplified if we use a polynomial approximation to the square root of the test information function, $[I(\theta)]^{1/2}$, in the least squares sense which is accomplished by using the method of moments (cf. Samejima and Livingston, 1979) for the meaningful interval of τ. Thus (2.9) can be changed to the form

$$\begin{aligned} \tau &\doteq C_1^{-1} \sum_{k=0}^{m} \alpha_k (k+1)^{-1} \theta^{k+1} + C_0 \\ &= \sum_{k=0}^{m+1} \alpha_k^* \theta^k \quad , \end{aligned} \tag{2.11}$$

where α_k $(k = 0, 1, \ldots, m)$ is the k-th coefficient of the polynomial of degree m approximating the square root of $I(\theta)$, and α_k^* is the new k-th coefficient which is given by

$$\alpha_k^* \begin{cases} = C_0 & k = 0 \\ = (C_1 k)^{-1} \alpha_{k-1} & k = 1, 2, \ldots, m+1 \end{cases} . \tag{2.12}$$

With this transformation of θ to τ and by virtue of (2.10), we can use the asymptotic normality with the two parameters, τ and C_1^{-1}, as the approximation to the conditional distribution of the maximum likelihood estimator $\hat{\tau}$, given its true value τ (cf. Samejima, 1981b). Then the first through fourth conditional moments of τ, given $\hat{\tau}$, can be obtained from the density function, $g^*(\hat{\tau})$, of $\hat{\tau}$ and from the constant C_1 by the following four formulae (cf. Samejima, 1981b):

$$E(\tau \mid \hat{\tau}) = \hat{\tau} + C_1^{-2} \frac{d}{d\hat{\tau}} \log g^*(\hat{\tau}) , \tag{2.13}$$

$$Var.(\tau \mid \hat{\tau}) = C_1^{-2}[1 + C_1^{-2} \frac{d^2}{d\hat{\tau}^2} \log g^*(\hat{\tau})] , \tag{2.14}$$

$$E[\{\tau - E(\tau \mid \hat{\tau})\}^3 \mid \hat{\tau}] = C_1^{-6}[\frac{d^3}{d\hat{\tau}^3} \log g^*(\hat{\tau})] \tag{2.15}$$

and

$$\begin{aligned} E[\{\tau - E(\tau \mid \hat{\tau})\}^4 \mid \hat{\tau}] &= C_1^{-4}[3 + 6C_1^{-2}\{\frac{d^2}{d\hat{\tau}^2} \log g^*(\hat{\tau})\} + 3C_1^{-4}\{\frac{d^2}{d\hat{\tau}^2} \log g^*(\hat{\tau})\}^2 \\ &\quad + C_1^{-4}\{\frac{d^4}{d\hat{\tau}^4} \log g^*(\hat{\tau})\}] . \end{aligned} \tag{2.16}$$

This density function, $g^*(\hat{\tau})$, can be estimated by fitting a polynomial, using the method of moments (cf. Samejima and Livingston, 1979), as we did in the transformation of θ to τ, based upon the empirical set of $\hat{\tau}$'s. Note that in the above formulae the first moment is about the origin, while the other three are about the mean.

The two coefficients, β_1 and β_2, and Pearson's criterion κ are obtained by

$$\beta_1 = \mu_3^2 \mu_2^{-3} , \tag{2.17}$$

$$\beta_2 = \mu_4 \mu_2^{-2} \tag{2.18}$$

and

$$\kappa = \beta_1(\beta_2 + 3)^2[4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)]^{-1} , \tag{2.19}$$

by substituting μ_2, μ_3 and μ_4 by $Var.(\tau \mid \hat{\tau})$, $E[\{\tau - E(\tau \mid \hat{\tau})\}^3 \mid \hat{\tau}]$ and $E[\{\tau - E(\tau \mid \hat{\tau})\}^4 \mid \hat{\tau}]$ respectively, which are obtained by formulae (2.14), (2.15) and (2.16).

In the Bivariate P.D.F. Approach, we approximate the bivariate distribution of the transformed latent trait τ and its maximum likelihood estimate $\hat{\tau}$ *for each subpopulation of examinees who share*

the same discrete item response to a specified item. Thus the procedure must be repeated as many times as the number of discrete item response categories for each separate item. It is rather a time-consuming approach, and the CPU time for the item calibration increases almost proportionally to the number of *new* items.

In contrast to this, Conditional P.D.F. Approach deals with the total population of subjects, and all the items together. Effort is focused upon the approximation of the conditional distribution of r, given θ, for the total population of examinees, and then the result is branched into separate discrete item response subpopulations for each item.

If we compare the two approaches with each other, therefore, we can say that Bivariate P.D.F. Approach is an orthodox approach, while Conditional P.D.F. Approach needs an assumption that the conditional distribution of r, given θ, is unaffected by the different subpopulations of examinees. While this assumption can only be tolerated in most cases, the latter approach has two big advantages in the sense that the CPU time required in item calibration is substantially less, and that it does not have to deal with subgroups of small numbers of subjects in approximating the joint bivariate distributions of r and θ.

In each of these two approaches, we can choose one of the four methods listed earlier in estimating the bivariate density of r and θ, or the conditional density of r, given its maximum likelihood estimate $\hat{\theta}$. In so doing, in the Pearson System Method, we use all four conditional moments of r, given θ, which are estimated through the formulae (2.13) through (2.16), and, using Pearson's criterion κ, which is given by (2.19), one of the Pearson System density functions is selected. In the Two-Parameter Beta Method two of the four parameters of the Beta density function, i.e., the lower and upper endpoints of the interval of r for which the Beta density is positive, are a priori given, and the other two parameters are estimated by using the first two conditional moments of r, given $\hat{\theta}$, which are provided by (2.13) and (2.14), respectively. In the Normal Approach Method, again we use only the first two conditional moments of r, given θ, as the first and second parameters of the normal density function.

If we compare these three methods, it will be appropriate to say that both Two-Parameter Beta Method and Normal Approach Method are simpler versions of Pearson System Method. And yet the latter two methods have an advantage of using only the first two estimated conditional moments of r, given θ, whereas the former requires the additional third and fourth conditional moments, whose estimations are less accurate compared with those of the first two conditional moments. If we compare the Two-Parameter Beta Method with the Normal Approach Method, we will notice that the former allows non-symmetric density functions, while the latter does not. This is an advantage of the Two-Parameter Beta Method over the Normal Approach Method, and yet the former has the disadvantage of the requirement that two of the four parameters should a priori be set.

Lognormal Approach Method was developed later, which uses up to the third conditional moment and allows more flexibilities in the shape of the conditional distribution of r, given $\hat{\theta}$, than the Normal Approach Method. It was intended that a *happy medium* between the Pearson System Method and the Normal Approach Method would be realized, in the effort of ameliorating the disadvantages of these two methods and of keeping their separate advantages.

[II.4] Possible Non-Monotonicities of the Operating Characteristics

As early as in 1968 the author wrote about and discussed the conceivable non-monotonicity of the operating characteristic of the correct answer of the multiple-choice test item, which is based strictly upon theory (cf. Samejima, 1968). Since then, such a phenomenon has actually been observed with empirical data. For example, Lord and Novick reported such a curve when they plotted the percent of the correct answer against the test score for each item as an approximation to the item characteristic function (cf. Lord and Novick, 1968, Chapter 16). Since, as their Theorem 16.4.1 states, *the average,*

over all items, of the sample item-test regressions falls along a straight line through the origin with forty-five degree slope, such a *dip* cannot be detected for an easy item even if it exists, as far as we use the item-test regression as an approximation. It is quite possible, therefore, that there are more than one item among those items that have such *dips*; only they were not detected.

In the past years various sets of data based upon the Vocabulary Subtest of the Iowa Tests of Basic Skills, upon Shiba's Word/Phrase Comprehension Tests, ASVAB Tests of Word Knowledge and of Math Knowledge, etc., have been analysed by using, mainly, the Simple Sum Procedure of the Conditional P.D.F. Approach combined with the Normal Approach Method (cf. Samejima, 1981b).These tests consist of multiple-choice test items, with four or five alternative answers in each item. As the result, we have discovered non-monotonic operating characteristics of the correct answer for some of the items, as well as differential information coming from the estimated operating characteristics of the incorrect alternative answers, which are called *plausibility functions*.

Such *discoveries* of non-monotonic operating characteristics can best be accomplished by using a nonparametric approach to the estimation of the operating characteristics. After the operating characteristics have been discovered by using the nonparametric approach, however, it may be wise to search for mathematical models that fit the results, and to estimate item parameters accordingly, so that we shall be able to take advantage of the mathematical simplicity coming from the parameterisation.

References

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

[2] Samejima, F. Application of the graded response model to the nominal response and multiple-choice situations. *UNC Psychometric Laboratory Report, 63*, 1968.

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

[4] Samejima, F. Effects of individual optimisation in setting boundaries of dichotomous items on accuracy of estimation. *Applied Psychological Measurement, 1*, 1977a, 77-94.

[5] Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1, 1977b, 233-247.

[6] Samejima, F. Estimation of the operating characteristics of item response categories I: Introduction to the Two-Parameter Beta Method. *ONR/RR-77-1*, 1977c.

[7] Samejima, F. Estimation of the operating characteristics of item response categories II: Further development of the Two-Parameter Beta Method. *ONR/RR-78-1*, 1978a.

[8] Samejima, F. Estimation of the operating characteristics of item response categories III: The Normal Approach Method and the Pearson System Method. *ONR/RR-78-2*, 1978b.

[9] Samejima, F. Estimation of the operating characteristics of item response categories IV: Comparison of the different methods. *ONR/RR-78-3*, 1978c.

[10] Samejima, F. Estimation of the operating characteristics of item response categories V: Weighted Sum Procedure in the Conditional P.D.F. Approach. *ONR/RR-78-4*, 1978d.

[11] Samejima, F. Estimation of the operating characteristics of item response categories VI: Proportioned Sum Procedure in the Conditional P.D.F. Approach. *ONR/RR-78-5*, 1978e.

[12] Samejima, F. Estimation of the operating characteristics of item response categories VII: Bivariate P.D.F. Approach with Normal Approach Method. *ONR/RR-78-6*, 1978f.

[13] Samejima, F. Estimation of the operating characteristics when the test information of the Old Test is not constant I: Rationale. *ONR/RR-80-2*, 1980a.

[14] Samejima, F. Estimation of the operating characteristics when the test information of the Old Test is not constant II: Simple Sum Procedure of the Conditional P.D.F. Approach/Normal Approach Method using three subtests of the Old Test. *ONR/RR-80-4*, 1980b.

[15] Samejima, F. Estimation of the operating characteristics when the test information of the Old Test is not constant II: Simple Sum Procedure of the Conditional P.D.F. Approach/Normal Approach Method using three subtests of the Old Test, No. 2. *ONR/RR-81-2*, 1981a.

[16] Samejima, F. Final Report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item. *Final Report of N00014-77-C-0360*, Office of Naval Research, 1981b.

[17] Samejima, F. Final Report: Advancement of latent trait theory. *Final Report of N00014-81-C-0569*, Office of Naval Research, 1988.

[18] Samejima, F. and Changas, P. S. How small the number of test items can be for the basis of estimating the operating characteristics of the discrete responses to unknown test items. *ONR/RR-81-3*, 1981.

[19] Samejima, F. and Livingston, P. S. Method of moments as the least squares solution for fitting a polynomial. *ONR/RR-79-2*, 1979.

III Proposal of Two Modification Formulae of the Test Information Function

Although the reciprocal of the test information function $I(\theta)$ provides us with a minimum variance bound for any *unbiased* estimator of θ (cf. Kendall and Stuart, 1961), since the maximum likelihood estimate, which is denoted by $\hat{\theta}_V$, is only *asymptotically unbiased*, for a finite number of items we need to examine if the bias of $\hat{\theta}_V$ of a given test over the meaningful range of θ is practically nil, before we consider this reciprocal as a minimum variance bound. It has been shown (Samejima, 1977a, 1977b) that in many cases the conditional distribution of $\hat{\theta}_V$, given θ, converges to $N(\theta, [I(\theta)]^{-1/2})$ relatively quickly. On the other hand, we have also noticed that the speed of convergence is not the same even if the amount of test information is kept equal. This has been demonstrated by using Constant Information Model (Samejima, 1979a), which is represented by

$$P_g(\theta) = \sin^2[a_g(\theta - b_g) + (\pi/4)] \quad , \tag{3.1}$$

where, as before, $P_g(\theta)$ denotes the operating characteristic of the correct answer, and a_g (> 0) and b_g are the item discrimination and difficulty parameters, respectively. This model provides us with a constant amount of item information $I_g(\theta)$ which equals $4a_g^2$ for the interval of θ,

$$-\pi[4a_g]^{-1} + b_g < \theta < \pi[4a_g]^{-1} + b_g \tag{3.2}$$

(cf. Samejima, 1979b).

Thus two modification formulae of the test information function $I(\theta)$ have been proposed in the present research in order to provide better measures of local accuracies of the estimation of θ, when the maximum likelihood estimation is used. They start from the search for a minimum variance bound, and from a minimum bound of the mean squared error, of *any* estimator, biased or unbiased.

[III.1] Minimum Variance Bound

Let θ^*_V denote any estimator of θ. We can write in general

$$E(\theta^*_V \mid \theta) = \theta + E[(\theta^*_V - \theta) \mid \theta] \quad . \tag{3.3}$$

When the item responses are discrete, we have

$$E(\theta^*_V \mid \theta) = \sum_V \theta^*_V \; L_V(\theta) = \sum_V \theta^*_V \; P_V(\theta) \quad , \tag{3.4}$$

where $L_V(\theta)$ denotes the likelihood function. Differentiating both sides of (3.4) with respect to θ, we obtain

$$\begin{aligned} \frac{\partial}{\partial\theta} E(\theta^*_V \mid \theta) &= \frac{\partial}{\partial\theta} [\sum_V \theta^*_V \; P_V(\theta)] = \sum_V \theta^*_V [\frac{\partial}{\partial\theta} P_V(\theta)] \\ &= \sum_V [\theta^*_V - E(\theta^*_V \mid \theta)] \, [\frac{\partial}{\partial\theta} P_V(\theta)] \, . \end{aligned} \tag{3.5}$$

We can write

$$\frac{\partial}{\partial\theta} P_V(\theta) = [\frac{\partial}{\partial\theta} \log P_V(\theta)] \; P_V(\theta) \quad , \tag{3.6}$$

and using this we can rewrite (3.5) into the form

$$\frac{\partial}{\partial\theta} E(\theta^*_V \mid \theta) = \sum_V [\theta^*_V - E(\theta^*_V \mid \theta)] \, [\frac{\partial}{\partial\theta} \log P_V(\theta)] \; P_V(\theta) \quad . \tag{3.7}$$

From this result, by the Cramér-Rao inequality, we obtain

$$[\frac{\partial}{\partial\theta} E(\theta^*_V \mid \theta)]^2 \leq Var.(\theta^*_V \mid \theta) \; E[\{\frac{\partial}{\partial\theta} \log P_V(\theta)\}^2 \mid \theta] \quad . \tag{3.8}$$

Since we can write

$$E[\{\frac{\partial}{\partial\theta} \log L_V(\theta)\}^2 \mid \theta] = -E[\frac{\partial^2}{\partial\theta^2} \log L_V(\theta) \mid \theta] \quad , \tag{3.9}$$

from this, (2.7), (2.8) and (3.3) we can rewrite and rearrange the inequality (3.8) into the form

$$Var.(\theta^*_V \mid \theta) \geq [\frac{\partial}{\partial\theta} E(\theta^*_V \mid \theta)]^2 \; [I(\theta)]^{-1} = [1 + \frac{\partial}{\partial\theta} E(\theta^*_V - \theta \mid \theta)]^2 \; [I(\theta)]^{-1} \quad , \tag{3.10}$$

whose rightest hand side provides us with the minimum variance bound of the conditional distribution of any estimator θ^*_V. When θ^*_V is biased, the size of the minimum variance bound is determined by the second term of the first factor of the minimum bound, and the result can be greater or less than the reciprocal of the test information function depending upon the sign of this partial derivative.

[III.2] First Modified Test Information Function

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

$$P_g(\theta) = c_g + (1 - c_g)[1 + \exp\{-Da_g(\theta - b_g)\}]^{-1} \quad , \tag{3.11}$$

where a_g, b_g, and c_g 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 $B(\hat{\theta}_V \mid \theta)$ can be written as

$$B(\hat{\theta}_V \mid \theta) = D[I(\theta)]^{-2} \sum_{g=1}^{n} a_g I_g(\theta)[\psi_g(\theta) - \frac{1}{2}] \quad , \tag{3.12}$$

where

$$\psi_g(\theta) = [1 + \exp\{-Da_g(\theta - b_g)\}]^{-1} \tag{3.13}$$

(cf. Lord, 1983). We can see in the above formula of the MLE bias function that the bias should be negative when $\psi_g(\theta)$ is less than 0.5 for all the items, which is necessarily the case for lower values of θ, and should be positive when $\psi_g(\theta)$ is greater than 0.5 for all the items, i.e., for higher values of θ, and in between the bias tends to be close to zero, for the last factor in the formula assumes negative values for some items and positive values for some others, provided that the difficulty parameter b_g distributes widely.

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

$$\begin{aligned} B(\hat{\theta}_V \mid \theta) = E[\hat{\theta}_V - \theta \mid \theta] &= -(1/2)[I(\theta)]^{-2} \sum_{g=1}^{n} \sum_{k_g} A_{k_g}(\theta) P''_{k_g}(\theta) \\ &= -(1/2)[I(\theta)]^{-2} \sum_{g=1}^{n} \sum_{k_g} P'_{k_g}(\theta) P''_{k_g}(\theta)[P_{k_g}(\theta)]^{-1} \quad , \end{aligned} \tag{3.14}$$

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 θ, 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 (cf. Samejima, 1969, 1972). On the dichotomous response level, it can be reduced to the form

$$B(\hat{\theta}_V \mid \theta) = E[\hat{\theta}_V - \theta \mid \theta] = (-1/2)[I(\theta)]^{-2} \sum_{g=1}^{n} I_g(\theta) P''_g(\theta)[P'_g(\theta)]^{-1} \quad , \tag{3.15}$$

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

We can rewrite the inequality (3.10) for the maximum likelihood estimate $\hat{\theta}_V$

(3.16) $$Var.(\hat{\theta}_V \mid \theta) \geq [1 + \frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta)]^2 \, [I(\theta)]^{-1} \ .$$

Taking the reciprocal of the right hand side of (3.16), which is an approximate minimum variance bound of the maximum likelihood estimator, a modified test information function, $\Upsilon(\theta)$, is proposed by

(3.17) $$\Upsilon(\theta) = I(\theta) \, [1 + \frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta)]^{-2} \ .$$

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. 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 unbiased, then these two functions will assume the same value. We can write from (3.14) for the general form of the derivative of the MLE bias function

(3.18) $$\frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta) = \{I(\theta)\}^{-1}[(1/2)\{I(\theta)\}^{-1} \sum_{g=1}^{n} \sum_{k_g} (I_{k_g}(\theta) P''_{k_g}(\theta) - P'_{k_g}(\theta) P'''_{k_g}(\theta)\{P_{k_g}(\theta)\}^{-1}) - 2B(\hat{\theta}_V \mid \theta) I'(\theta)] \ ,$$

where $P'''_{k_g}(\theta)$ and $I'(\theta)$ denote the third and the first derivatives of $P_{k_g}(\theta)$ and $I(\theta)$ with respect to θ, respectively. It is obvious from (2.3) and (2.8) that we have

(3.19) $$I'_g(\theta) = \sum_{k_g} P'_{k_g}(\theta)[P''_{k_g}(\theta)\{P_{k_g}(\theta)\}^{-1} - I_{k_g}(\theta)]$$

and

(3.20) $$I'(\theta) = \sum_{g=1}^{n} I'_g(\theta) = \sum_{g=1}^{n} \sum_{k_g} P'_{k_g}(\theta)[P''_{k_g}(\theta)\{P_{k_g}(\theta)\}^{-1} - I_{k_g}(\theta)] \ ,$$

where $I'_g(\theta)$ is the first derivative of the item information function $I_g(\theta)$ with respect to θ. For a set of dichotomous items (3.18) becomes simplified into the form

(3.21) $$\frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta) = \{I(\theta)\}^{-1}[(1/2)\{I(\theta)\}^{-1} \sum_{g=1}^{n} \{P_g(\theta)\}^{-2}\{1 - P_g(\theta)\}^{-2}$$
$$(\{1 - 2P_g(\theta)\}\{P'_g(\theta)\}^2 P''_g(\theta) - P_g(\theta)\{1 - P_g(\theta)\}(\{P''_g(\theta)\}^2 + P'_g(\theta) P'''_g(\theta)))$$
$$- 2B(\hat{\theta}_V \mid \theta) \, I'(\theta)] \ ,$$

where $B(\hat{\theta}_V \mid \theta)$ is given by (3.15).

[III.3] Minimum Bound of the Mean Squared Error

When the estimator θ^*_V is conditionally *biased*, however small the conditional variance may be, it does not reflect the accuracy of estimation of θ. Thus the *mean squared error*, $E[(\theta^*_V - \theta)^2 \mid \theta]$, becomes a more important indicator of the accuracy. We can write for the mean squared error

(3.22) $$E[(\theta_V^* - \theta)^2 \mid \theta] = Var.(\theta_V^* \mid \theta) + [E(\theta_V^* \mid \theta) - \theta]^2$$

(cf. Kendall and Stuart, 1961). We can see in this formula that the mean squared error equals the conditional variance if θ_V^* is unbiased, and is greater than the variance when θ_V^* is biased. From this and the inequality (3.10) we obtain for the minimum bound of the mean squared error

(3.23) $$E[(\theta_V^* - \theta)^2 \mid \theta] \geq [1 + \frac{\partial}{\partial\theta} E(\theta_V^* - \theta \mid \theta)]^2 \, [I(\theta)]^{-1} + [E(\theta_V^* \mid \theta) - \theta]^2 \quad .$$

Note that this inequality holds for any estimator, θ_V^*, of θ.

[III.4] Second Modified Test Information Function

For the maximum likelihood estimate $\hat{\theta}_V$, we can rewrite the inequality (3.23) by using the MLE bias function, which is given by (3.14), to obtain

(3.24) $$E[(\hat{\theta}_V - \theta)^2 \mid \theta] \geq [1 + \frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta)]^2 \, [I(\theta)]^{-1} + [B(\hat{\theta}_V \mid \theta)]^2 \quad .$$

Taking the reciprocal of the right hand side of (3.24), which is an approximate minimum bound of the mean squared error of the maximum likelihood estimator, the second modified test information function, $\Xi(\theta)$, is proposed by

(3.25) $$\Xi(\theta) = I(\theta) \, \{[1 + \frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta)]^2 + I(\theta) \, [B(\hat{\theta}_V \mid \theta)]^2\}^{-1} \quad .$$

We can see that the difference between the two modification formulae of the test information function, which are defined by (3.17) and (3.25), respectively, is the second and last term in the braces of the right hand side of the formula (3.25). Since this term is nonnegative, there is a relationship

(3.26) $$\Xi(\theta) \leq \Upsilon(\theta) \quad ,$$

throughout the whole range of θ, regardless of the slope of the MLE bias function. If there is a range of θ where the maximum likelihood estimate is *unbiased,* then we will have for that range of θ

(3.27) $$\Xi(\theta) = \Upsilon(\theta) = I(\theta) \quad .$$

Since under a general condition the maximum likelihood estimator $\hat{\theta}_V$ is asymptotically unbiased, as the number of items approaches positive infinity, (3.27) holds *asymptotically* for all θ.

[III.5] Examples

Samejima has applied formula (3.15) for the MLE bias functions of the Iowa Level 11 Vocabulary Subtest and Shiba's Test J1 of Word/Phrase Comprehension, based upon the set of data collected for 2,356 and 2,259 subjects, respectively. These tests have forty-three and fifty-five dichotomously scored items, respectively, and following the normal ogive model, whose operating characteristic for the correct answer is given by

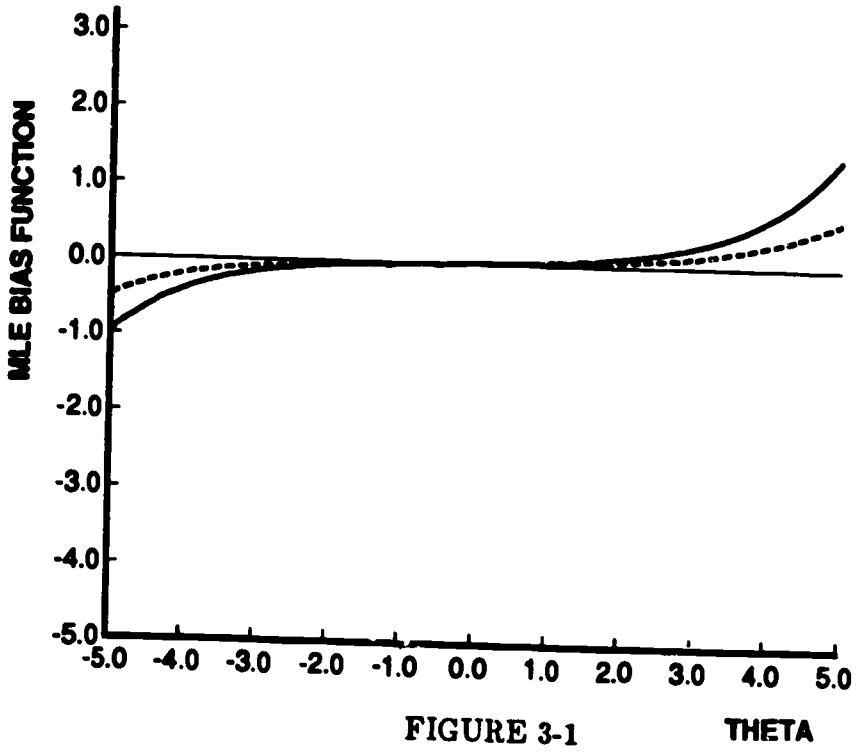

FIGURE 3-1

MLE Bias Functions of the Iowa Level 11 Vocabulary Subtest (Solid Line) and of Shiba's Test J1 of Word/Phrase Comprehension (Dashed Line), Following the Normal Ogive Model.

(3.28) $$P_g(\theta) = [2\pi]^{-1/2} \int_{-\infty}^{a_g(\theta - b_g)} e^{-u^2/2}\, du \ ,$$

the discrimination and difficulty parameters were estimated (Samejima, 1984a, 1984b). The resulting MLE bias functions are illustrated in Figure 3-1. We can see that in each of these two examples there is a wide range of θ, i.e., approximately (-2.0, 1.5), for which the maximum likelihood estimate of θ is practically unbiased. The amount of bias is especially small for Shiba's Test J1. Although this feature indicates good qualities of these tests, we still have to expect some biases when these tests are administered to groups of examinees whose ability distributes on the relatively lower side or on the relatively higher side of the ability scale.

When the MLE bias function of the test is *monotone increasing*, as are those illustrated in Figure 3-1, it is obvious from (3.17) that $\Upsilon(\theta)$ will assume lesser values than those of the original test information function $I(\theta)$ for lower and higher levels of θ, while these two functions are practically identical in between. The same applies to $\Xi(\theta)$, and we have the relationship,

(3.29) $$\Xi(\theta) \leq \Upsilon(\theta) \leq I(\theta) \ ,$$

throughout the whole range of θ.

In the normal ogive model, differentiating (3.28) twice with respect to θ and rearranging, we obtain

(3.30) $$P'_g(\theta) = [2\pi]^{-1/2} a_g \ \exp[-(1/2)\ a_g^2(\theta - b_g)^2]$$

and

(3.31) $$P_g''(\theta) = -a_g^2(\theta - b_g)\ P_g'(\theta) \quad .$$

Substituting (3.30) and (3.31) into (3.15) and rearranging, we can write for the MLE bias function following the normal ogive model on the dichotomous response level

(3.32) $$B(\hat{\theta}_V \mid \theta) = (1/2)\ [I(\theta)]^{-2} \sum_{g=1}^{n} a_g^2(\theta - b_g)\ I_g(\theta) \quad .$$

Differentiating (3.32) with respect to θ, we obtain

(3.33) $$\begin{aligned} \frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta) \quad = \quad & [I(\theta)]^{-2}[(1/2) \sum_{g=1}^{n} a_g^2\ [I_g'(\theta)(\theta - b_g) + I_g(\theta)] \\ & -[I(\theta)]^{-1}\ I'(\theta) \sum_{g=1}^{n} a_g^2(\theta - b_g)\ I_g(\theta)] \ . \end{aligned}$$

It is obvious from (2.4), (2.8) and (3.31) that we have

(3.34) $$I_g'(\theta) = I_g(\theta)\ [P_g'(\theta)\ \{2P_g(\theta) - 1\}\ (P_g(\theta)\{1 - P_g(\theta)\})^{-1} - 2a_g^2(\theta - b_g)]$$

and

(3.35) $$I'(\theta) = \sum_{g=1}^{n} I_g(\theta)\ [P_g'(\theta)\ \{2P_g(\theta) - 1\}\ (P_g(\theta)\{1 - P_g(\theta)\})^{-1} - 2a_g^2(\theta - b_g)] \quad .$$

Figure 3-2 shows the square roots of the original and the two modified test information functions for the Iowa Level 11 Vocabulary Subtest and for Shiba's Test J1 of Word/Phrase Comprehension, following the normal ogive model. In each of these figures, the curves respresenting the results of the two modification formulae assume lower values than the square root of the original test information function for all θ, as was expected from the shape of the MLE bias function in Figure 3-1. The discrepancies between the results of the two modification formulae are small, however, in each figure.

In the three-parameter logistic model, the operating characteristic of the correct answer is given by the formula (3.11), and Lord's MLE bias function for the three-parameter logistic model, which is given by (3.12), is readily applicable. Differentiating (3.11) three times with respec. to θ and rearranging, we can write

(3.36) $$P_g'(\theta) \ = (1 - c_g)\ Da_g\ \psi_g(\theta)\ [1 - \psi_g(\theta)] \quad ,$$

(3.37) $$P_g''(\theta) \ = (1 - c_g)\ D^2a_g^2\ \psi_g(\theta)\ [1 - \psi_g(\theta)]\ [1 - 2\psi_g(\theta)] = Da_g\ P_g'(\theta)[1 - 2\psi_g(\theta)]$$

and

(3.38) $$P_g'''(\theta) = D^2a_g^2\ P_g'(\theta)[1 - 6\psi_g(\theta) + 6\{\psi_g(\theta)\}^2] \quad ,$$

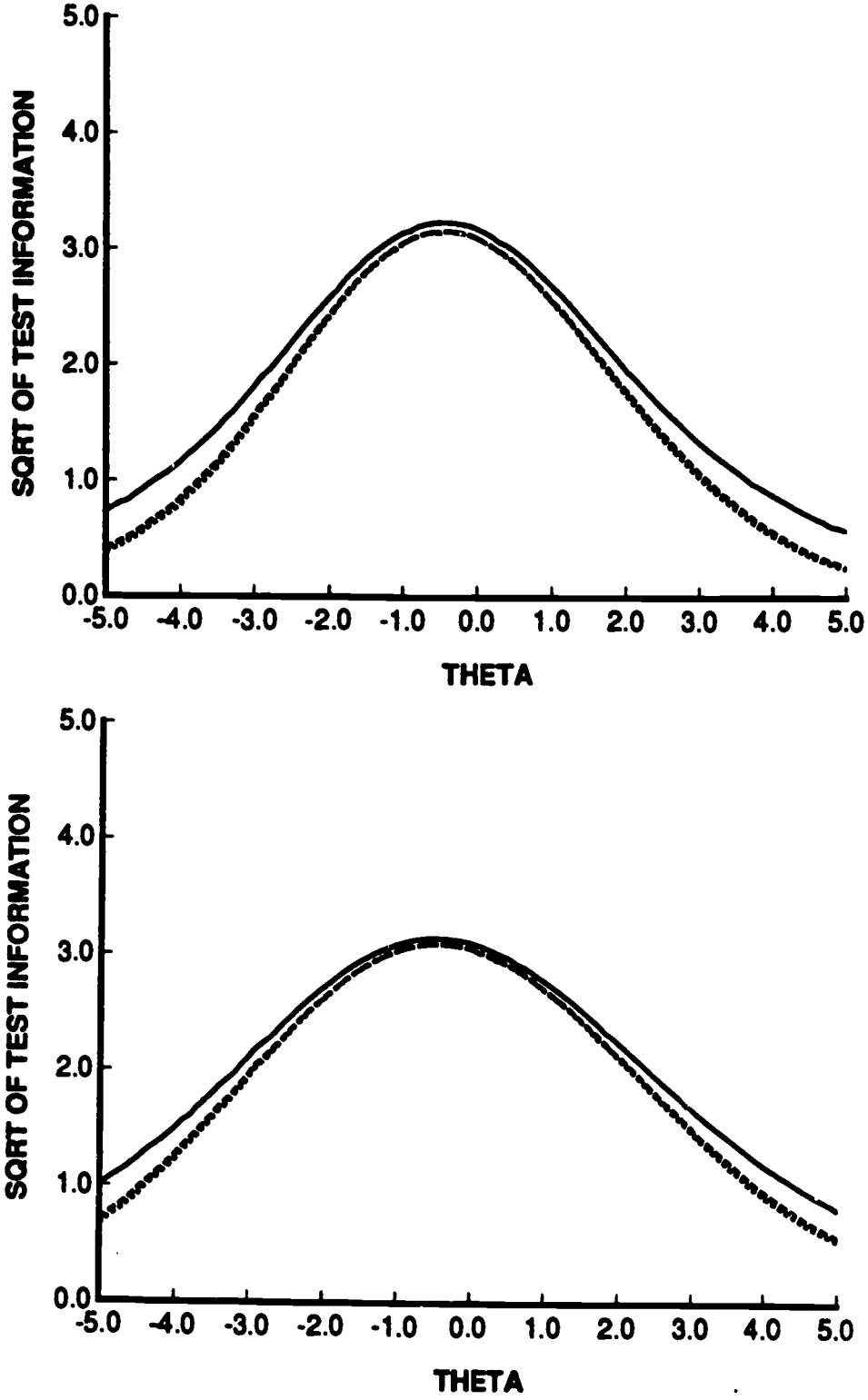

FIGURE 3-2

Square Roots of the Original (Solid Line) and the Two Modified (Dashed and Dotted Lines) Test Information Functions of the Iowa Level 11 Vocabulary Subtest, and Those of Shiba's Test J1 of Word/Phrase Comprehension, Following the Normal Ogive Model.

where $\psi_g(\theta)$ is defined by (3.13). Substituting (3.36) into (2.4) and rearranging, we obtain for the item information function

$$I_g(\theta) = (1-c_g)\, D^2 a_g^2 \{\psi_g(\theta)\}^2 [1-\psi_g(\theta)] [c_g + (1-c_g)\, \psi(\theta)]^{-1} \quad . \tag{3.39}$$

This and (2.8) will enable us to evaluate Lord's MLE bias function given by (3.12). Differentiating (3.12) with respect to θ and rearranging, we can write

$$\begin{aligned} \frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta) &= D\, \{I(\theta)\}^{-2} [\sum_{g=1}^{n} a_g\, I_g'(\theta) \{\psi_g(\theta) - (1/2)\} \\ &\quad + D \sum_{g=1}^{n} a_g^2\, I_g(\theta)\, \psi_g(\theta) \{1-\psi_g(\theta)\} \\ &\quad - 2\, I'(\theta)\, \{I(\theta)\}^{-1} \sum_{g=1}^{n} a_g\, I_g(\theta)\, \{\psi_g(\theta) - (1/2)\}] \, . \end{aligned} \tag{3.40}$$

We also obtain from (2.4), (3.11) and (2.8) the first derivatives of the item and the test information functions with respect to θ so that we have

$$\begin{aligned} I_g'(\theta) &= (1-c_g)\, D^3 a_g^3 \{\psi_g(\theta)\}^2 [1-\psi_g(\theta)] \{P_g(\theta)\}^{-1} \\ &\qquad [2 - 3\psi_g(\theta) - (1-c_g)\, \psi_g(\theta) \{1-\psi_g(\theta)\} \{P_g(\theta)\}^{-1}] \\ &= D a_g\, I_g(\theta) [2\{1-\psi_g(\theta)\} - \psi_g(\theta) \{P_g(\theta)\}^{-1}] \end{aligned} \tag{3.41}$$

and

$$I'(\theta) = D \sum_{g=1}^{n} a_g\, I_g(\theta) [2\{1-\psi_g(\theta)\} - \psi_g(\theta) \{P_g(\theta)\}^{-1}] \, , \tag{3.42}$$

and we can use these two results in (3.40) in order to evaluate $\frac{\partial}{\partial\theta} B(\hat{\theta}_V \mid \theta)$.

When $c_g = 0$, i.e., for the original logistic model on the dichotomous response level, these formulae become much more simplified, and we can write

$$P_g(\theta) = [1 + \exp\{-D a_g (\theta - b_g)\}]^{-1} = \psi_g(\theta) \quad , \tag{3.43}$$

$$P_g'(\theta) = D a_g\, \psi_g(\theta) [1-\psi_g(\theta)] \quad , \tag{3.44}$$

$$P_g''(\theta) = D^2 a_g^2\, \psi_g(\theta) [1-\psi_g(\theta)] [1-2\psi_g(\theta)] = D a_g\, P_g'(\theta) [1-2\psi_g(\theta)] \quad , \tag{3.45}$$

$$P_g'''(\theta) = D^3 a_g^3\, \psi_g(\theta) [1-\psi_g(\theta)] [1 - 6\psi_g(\theta) + 6\{\psi_g(\theta)\}^2] \quad , \tag{3.46}$$

$$I_g(\theta) = D^2 a_g^2\, \psi_g(\theta) [1-\psi_g(\theta)] \quad , \tag{3.47}$$

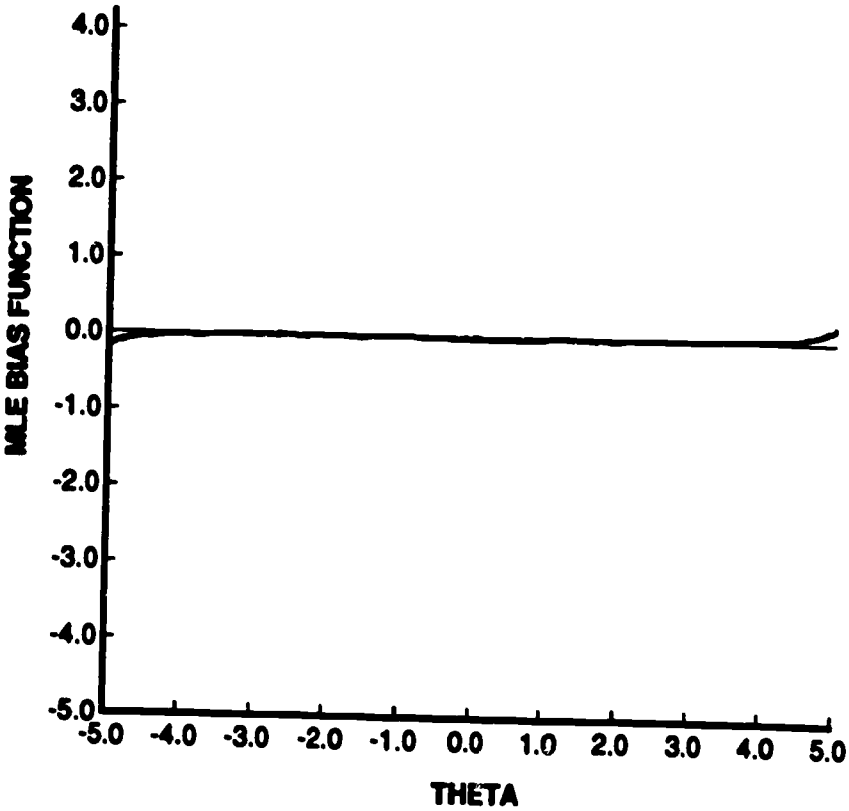

FIGURE 3-3

MLE Bias Functions of the Hypothetical Test of Thirty-Five Graded Test Items Following the Normal Ogive Model (Solid Line) and the Logistic Model (Dashed Line).

(3.48) $$I'_g(\theta) = D^3 a_g^3 \psi_g(\theta)[1-\psi_g(\theta)][1-2\psi_g(\theta)] = D\, a_g\, I_g(\theta)\,[1-2\psi_g(\theta)] ,$$

(3.49) $$I(\theta) = D^2 \sum_{g=1}^{n} a_g^2\, \psi_g(\theta)\,[1-\psi_g(\theta)]$$

and

(3.50) $$I'(\theta) = D \sum_{g=1}^{n} a_g\, I_g(\theta)\,[1-2\psi_g(\theta)] ,$$

respectively. Thus the two modified test information functions, $\Upsilon(\theta)$ and $\Xi(\theta)$, which are defined by (3.17) and (3.25), can be evaluated accordingly, both for the original logistic model and for the three-parameter logistic model.

The reader is directed to ONR/RR-90-1 (cf. Samejima, 1990) for the MLE bias functions and the square roots of the original and the two modified test information functions of the Iowa Level 11 Vocabulary Subtest and of Shiba's Test J1 of Word/Phrase Comprehension, following the logistic model by using the same sets of estimated item parameters and by setting $D = 1.7$. These results are similar to those following the normal ogive model, which are presented by Figures 3-1 and 3-2, except that

the square roots of the original and the modified test information functions are a little steeper, the characteristic of the logistic model in comparison with the normal ogive model.

In the homogeneous case of the graded response level (Samejima, 1969, 1972), the general formula for the operating characteristic of the item score x_g $(= 0, 1, ..., m_g)$ is given by

$$P_{x_g}(\theta) = P^*_{x_g}(\theta) - P^*_{x_g+1}(\theta) \ , \tag{3.51}$$

where

$$P^*_{x_g}(\theta) = \int_{-\infty}^{a_g(\theta - b_{x_g})} \phi_g(t)\, dt \ , \tag{3.52}$$

$$-\infty = b_0 < b_1 < b_2 < ... < b_{m_g} < b_{m_g+1} = \infty \ , \tag{3.53}$$

and $\phi_g(t)$ is some specified density function. When we replace the right hand side of (3.52) by that of (3.28) with b_g replaced by b_{x_g} and use the result in (3.51), we have the operating characteristic of x_g in the normal ogive model on the graded response level; when we do the same thing using the right hand side of (3.13), we obtain the operating characteristic of x_g in the logistic model on the graded response level.

A hypothetical test of thirty-five graded items, with three graded score categories each, which gives an approximately constant amount of test information for the interval of θ, (-3, 3), has been used repeatedly in the author's research (cf. Samejima, 1981, 1988). Figure 3-3 presents the MLE bias functions for this hypothetical test, following the normal ogive model and the logistic model on the graded response level, respectively. We can see that a practical unbiasedness holds for a very wide range of θ in both cases, as is expected for a set of graded test items whose response difficulty levels are widely distributed, an advantage of graded responses over dichotomous responses. We also notice that these two MLE bias functions are almost indistinguishable from each other. Figure 3-4 presents the square roots of the original and the two modified test information functions of this hypothetical test of graded items, following the normal ogive model and the logistic model. As is expected, the differences among the three functions are small for a wide range of θ in both cases. It is interesting to note, however, that in these figures the square roots of the modified test information functions assume higher values than the square root of the original test information function at certain points of θ, and this tendency is especially conspicuous in the results of the logistic model. This comes from the fact that the MLE bias functions, which are presented in Figure 3-3 for both models, have tiny ups and downs, and they are not strictly increasing in θ.

In each of the examples given above, the difficulty parameters of these items in each test distribute widely over the range of θ of interest, and this fact is the main reason that the MLE bias function assumes relatively small values for a wide range of θ. We also notice that the resulting two modified test information functions are reasonably close to the original test information function.

For the sake of comparison, Figure 3-5 presents the MLE bias function and the square roots of the original and the two modified test information functions, for a hypothetical test of thirty equivalent, dichotomous items with the common item parameters, $a_g = 1.0$ and $b_g = 0.0$, following the logistic model. We can see in the first graph of Figure 3-5 that the amount of bias increases rapidly outside the range of θ, $(-1.0, 1.0)$. The resulting square roots of the two modified test information functions demonstrate substantially large decrements from the original $[I(\theta)]^{1/2}$ outside this interval of θ, as we can see in the second graph of Figure 3-5.

We also notice that in all these examples there are not substantial differences between the results of the two modification formulae. This indicates that in these examples it does not make so much

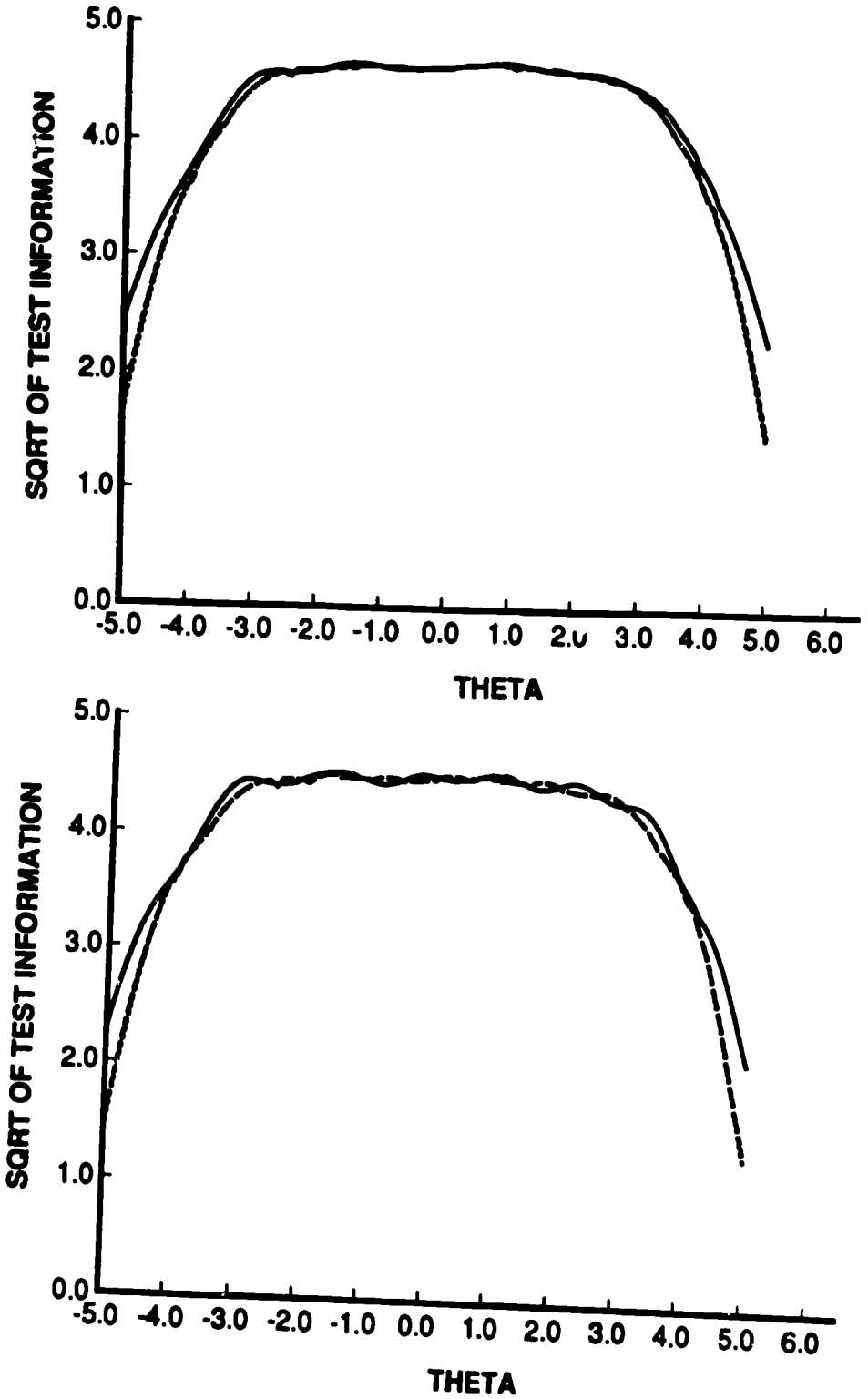

FIGURE 3-4

Square Roots of the Original (Solid Line) and the Two Modified (Dashed and Dotted Lines) Test Information Functions of the Hypothetical Test of Thirty-Five Graded Test Items Following the Normal Ogive Model and the Logistic Model, Respectively.

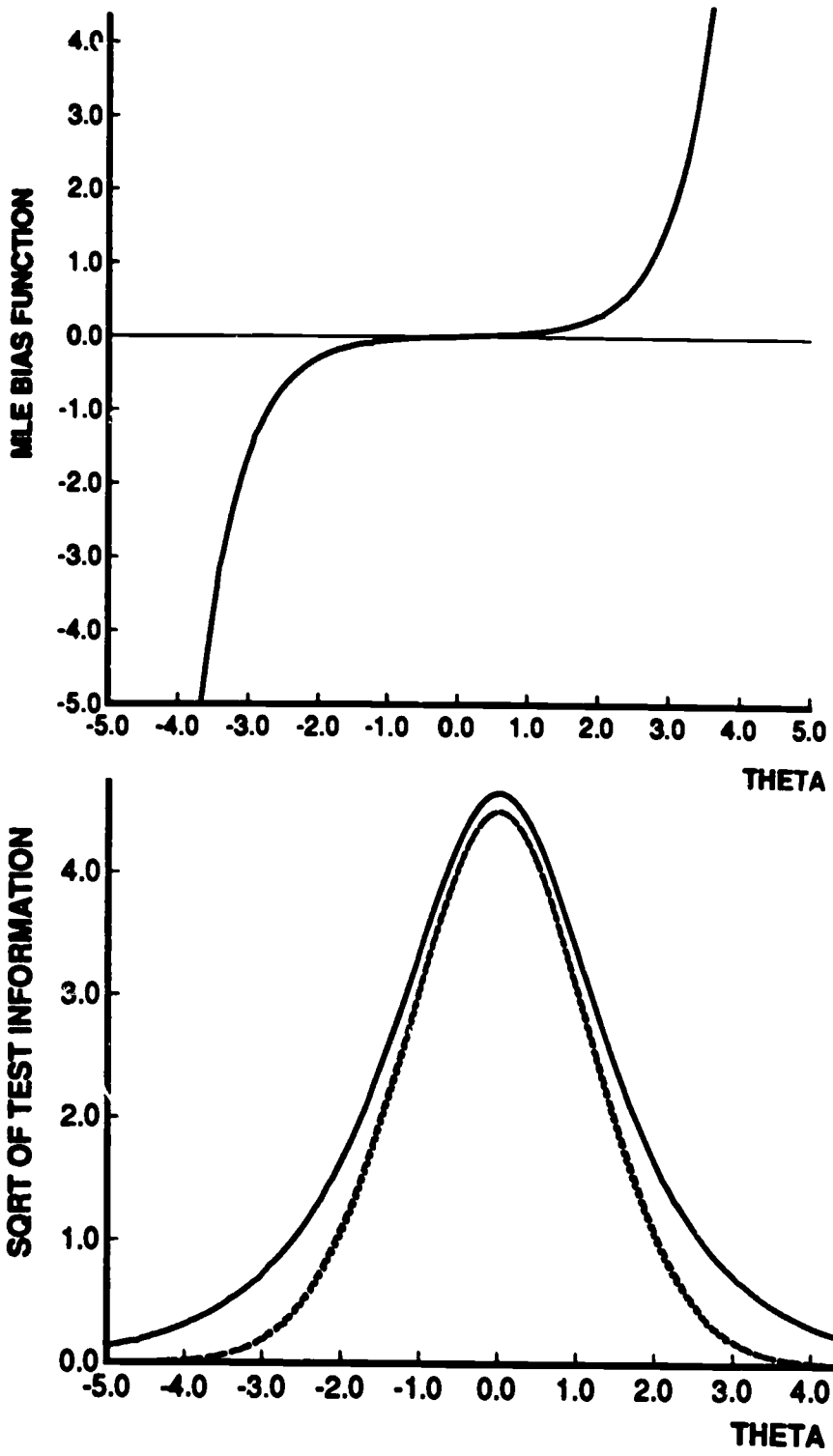

FIGURE 3-5

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 (Above), and Square Roots of the Original (Solid Line) and the Two Modified (Dashed and Dotted Lines) Test Information Functions of the Same Test (Below).

difference if we choose Modification Formula No. 1 or Modification Formula No. 2. We should not generalise this conclusion to other situations, however, until we have tried these modification formulae on different types of data sets.

[III.6] Minimum Bounds of Variance and Mean Squared Error for the Transformed Latent Variable

Since most psychological scales, including those in latent trait models, are subject to *monotone* transformation, we need to consider information functions that are based upon the transfomed latent variable. Let τ denote a transformed latent variable, i.e.,

$$\tau = \tau(\theta) \ . \tag{3.54}$$

We assume that τ is strictly increasing in, and three times differentiable with respect to, θ, and vice versa. We have for the operating characteristic, $P^*_{k_g}(\tau)$, of the discrete item response k_g, which is defined as a function of τ,

$$P^*_{k_g}(\tau) = prob.[k_g \mid \tau] = prob.[k_g \mid \theta] = P_{k_g}(\theta) \ , \tag{3.55}$$

and by local independence we can write for the operating characteristic of the response pattern, $P^*_V(\tau)$,

$$P^*_V(\tau) = \prod_{k_g \epsilon V} P^*_{k_g}(\tau) = \prod_{k_g \epsilon V} P_{k_g}(\theta) = P_V(\theta) \ . \tag{3.56}$$

As before, the item response information function, $I^*_{k_g}(\tau)$, is defined by

$$I^*_{k_g}(\tau) = -\frac{\partial^2}{\partial \tau^2} \log P^*_{k_g}(\tau) \ , \tag{3.57}$$

and for the item information function, $I^*_g(\tau)$, and the test information function, $I^*(\tau)$, we can write from (3.57), (2.3) and (2.8)

$$\begin{aligned} I^*_g(\tau) &= \sum_{k_g} I^*_{k_g}(\tau) \, P^*_{k_g}(\tau) = \sum_{k_g} [\frac{\partial}{\partial \tau} P^*_{k_g}(\tau)]^2 \, [P^*_{k_g}(\tau)]^{-1} \\ &= \sum_{k_g} [\frac{\partial}{\partial \theta} P_{k_g}(\theta) \, \frac{\partial \theta}{\partial \tau}]^2 \, [P_{k_g}(\theta)]^{-1} = I_g(\theta) \, [\frac{\partial \theta}{\partial \tau}]^2 \end{aligned} \tag{3.58}$$

and

$$I^*(\tau) = \sum_{g=1}^{n} I^*_g(\tau) = I(\theta) \, [\frac{\partial \theta}{\partial \tau}]^2 \ , \tag{3.59}$$

respectively. Let τ^*_V be any estimator of τ, which may be biased or unbiased. In general, we can write

$$E(\tau^*_V \mid \tau) = \tau + E(\tau^*_V - \tau \mid \tau) \ , \tag{3.60}$$

and, differentiating (3.60) with respect to θ, we obtain

(3.61) $$\frac{\partial}{\partial\theta}E(\tau_V^* \mid \tau) = \frac{\partial\tau}{\partial\theta} + \frac{\partial}{\partial\theta}E(\tau_V^* - \tau \mid \tau) \ .$$

Since from (3.56) we can also write for $E(\tau_V^* \mid \tau)$

(3.62) $$E(\tau_V^* \mid \tau) = \sum_V \tau_V^* \ P_V^*(\tau) = \sum_V \tau_V^* \ P_V(\theta) \ ,$$

differentiating (3.62) with respect to θ and following a logic similar to that used in Section 3.1, we obtain

(3.63) $$\begin{aligned}\frac{\partial}{\partial\theta}E(\tau_V^* \mid \tau) &= \frac{\partial}{\partial\theta}\sum_V \tau_V^* \ P_V(\theta) = \sum_V [\tau_V^* - E(\tau_V^* \mid \tau)] \ [\frac{\partial}{\partial\theta}P_V(\theta)] \\ &= \sum_V [\tau_V^* - E(\tau_V^* \mid \tau)] \ [\frac{\partial}{\partial\theta}\log P_V(\theta)] \ P_V(\theta) \ .\end{aligned}$$

By the Cramér-Rao inequality, we can write

(3.64) $$[\frac{\partial}{\partial\theta}E(\tau_V^* \mid \tau)]^2 \leq Var.(\tau_V^* \mid \tau) \ E[\{\frac{\partial}{\partial\theta}\log P_V(\theta)\}^2] \ ,$$

and from this, (2.7), (2.8), (3.10) and (3.61) we obtain

(3.65) $$\begin{aligned}Var.(\tau_V^* \mid \tau) &\geq [\frac{\partial}{\partial\theta}E(\tau_V^* \mid \tau)]^2 \ [I(\theta)]^{-1} \\ &= [\frac{\partial\tau}{\partial\theta} + \frac{\partial}{\partial\theta}E(\tau_V^* - \tau \mid \tau)]^2 \ [I(\theta)]^{-1} \ .\end{aligned}$$

Thus the rightest hand side of (3.65) provides us with the minimum variance bound of any estimator of τ. When τ_V^* is an unbiased estimator of τ, the second term of the first factor of the rightest hand side of (3.65) equals zero, and by virtue of (3.59) the inequality is reduced to

(3.66) $$Var.(\tau_V^* \mid \tau) \geq [\frac{\partial\tau}{\partial\theta}]^2 \ [I(\theta)]^{-1} = [I^*(\tau)]^{-1} \ .$$

For the mean squared error, $E[(\tau_V^* - \tau)^2 \mid \tau]$, we can write

(3.67) $$E[(\tau_V^* - \tau)^2 \mid \tau] = Var.(\tau_V^* \mid \tau) + [E(\tau_V^* \mid \tau) - \tau]^2 \ ,$$

and from this and (3.65) we obtain

(3.68) $$E[(\tau_V - \tau)^2 \mid \tau] \geq [\frac{\partial\tau}{\partial\theta} + \frac{\partial}{\partial\theta}E(\tau_V^* - \tau \mid \tau)]^2 \ [I(\theta)]^{-1} + [E(\tau_V^* \mid \tau) - \tau]^2 \ .$$

[III.7] Modified Test Information Functions Based upon the Transformed Latent Variable

The maximum likelihood estimator, $\hat{\tau}_V$, of τ, can be obtained by the direct transformation of the maximum likelihood estimate, $\hat{\theta}_V$, of θ, i.e.,

$$\hat{\tau}_V = \tau(\hat{\theta}_V) \quad . \tag{3.69}$$

Let $B^*(\hat{\tau}_V \mid \tau)$ be the MLE bias function defined for the transformed latent variable τ, i.e.,

$$B^*(\hat{\tau}_V \mid \tau) = E(\hat{\tau}_V - \tau \mid \tau) \quad . \tag{3.70}$$

From this, (3.65) and (3.68) we obtain

$$Var.(\hat{\tau}_V \mid \tau) \geq [\frac{\partial \tau}{\partial \theta} + \frac{\partial}{\partial \theta} B^*(\hat{\tau}_V \mid \tau)]^2 [I(\theta)]^{-1} \tag{3.71}$$

and

$$E[(\hat{\tau}_V - \tau)^2 \mid \tau] \geq [\frac{\partial \tau}{\partial \theta} + \frac{\partial}{\partial \theta} B^*(\hat{\tau}_V \mid \tau)]^2 [I(\theta)]^{-1} + [B^*(\hat{\tau}_V \mid \tau)]^2 \quad . \tag{3.72}$$

The reciprocals of the right hand sides of the above two inequalities provide us with the two modified test information functions for the transformed latent variable τ, i.e.,

$$\Upsilon^*(\tau) = I(\theta) \, [\frac{\partial \tau}{\partial \theta} + \frac{\partial}{\partial \theta} B^*(\hat{\tau}_V \mid \tau)]^{-2} \tag{3.73}$$

and

$$\Xi^*(\tau) = I(\theta) \, [\{\frac{\partial \tau}{\partial \theta} + \frac{\partial}{\partial \theta} B^*(\hat{\tau}_V \mid \tau)\}^2 + I(\theta) \, \{B^*(\hat{\tau}_V \mid \tau)\}^2]^{-1} \quad . \tag{3.74}$$

In the general case of discrete item responses we can write for the MLE bias function $B^*(\hat{\tau}_V \mid \tau)$ and its derivative with respect to θ

$$\begin{aligned} B^*(\hat{\tau}_V \mid \tau) &= B(\hat{\theta}_V \mid \theta)[\frac{\partial \theta}{\partial \tau}]^{-1} - (1/2)[I(\theta)]^{-1}[\frac{\partial \theta}{\partial \tau}]^{-3} \frac{\partial^2 \theta}{\partial \tau^2} \\ &= B(\hat{\theta}_V \mid \theta)\frac{\partial \tau}{\partial \theta} + (1/2)[I(\theta)]^{-1}\frac{\partial^2 \tau}{\partial \theta^2} \quad , \end{aligned} \tag{3.75}$$

and

$$\begin{aligned} \frac{\partial}{\partial \theta} B^*(\hat{\tau}_V \mid \tau) &= B(\hat{\theta}_V \mid \theta) \, \frac{\partial^2 \tau}{\partial \theta^2} + [\frac{\partial}{\partial \theta} B(\hat{\theta}_V \mid \theta)] \, \frac{\partial \tau}{\partial \theta} \\ &\quad + (1/2)[I(\theta)]^{-2}[I(\theta) \, \frac{\partial^3 \tau}{\partial \theta^3} - I'(\theta) \, \frac{\partial^2 \tau}{\partial \theta^2}] \quad , \end{aligned} \tag{3.76}$$

respectively (cf. Samejima, 1987). Thus we can use (3.75) and (3.76) in evaluating the modified test information functions, $\Upsilon^*(\tau)$ and $\Xi^*(\tau)$, which are given by (3.73) and (3.74).

[III.8] Discussion and Conclusions

A minimum bound of any estimator, biased or unbiased, has been considered, and, based on that, Modification Formula No. 1 has been proposed for the maximum likelihood estimator, in place of the test information function. A minimum bound of the mean squared error of any estimator has also been considered, and, based on that, Modification Formula No. 2 in the same context has been proposed. Examples have been given. These topics have also been discussed and observed for the monotonically transformed latent variable.

It is expected that these two modification formulae of the test information function can effectively be used in order to supplement a relative weakness of the test information function in certain situations. Results are yet to come.

References

[1] Kendall, M. G. and Stuart, A. *The advanced theory of statistics. Vol. 2.* New York: Hafner, 1961.

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

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

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

[5] Samejima, F. Effects of individual optimisation in setting boundaries of dichotomous items on accuracy of estimation. *Applied Psychological Measurement, 1*, 1977a, 77-94.

[6] Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1, 1977b, 233-247.

[7] Samejima, F. Constant information model: a new promising item characteristic function. *ONR/RR-79-1*, 1979a.

[8] Samejima, F. Convergence of the conditional distribution of the maximum likelihood estimate, given latent trait, to the asymptotic normality: Observations made through the constant information model. *ONR/RR-79-3*, 1979b.

[9] Samejima, F. Final Report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item. *Final Report of N00014-77-C-0360*, Office of Naval Research, 1981.

[10] Samejima, F. Plausibility functions of Iowa Vocabulary test items Estimated by the Simple Sum Procedure of the Conditional P.D.F. Approach. *ONR/RR-84-1*, 1984a.

[11] Samejima, F. Comparison of the estimated item parameters of Shiba's Word/Phrase Comprehension Tests obtained by Logist 5 and those by the tetrachoric method. *ONR/RR-84-2*, 1984b.

[12] Samejima, F. Bias function of the maximum likelihood estimate of ability for discrete item responses. *ONR/RR-87-1*, 1987.

[13] Samejima, F. Final Report: Advancement of latent trait theory. *Final Report of N00014-81-C-0569*, Office of Naval Research, 1988.

[14] Samejima, F. Modifications of the test information function. *ONR/RR-90-1*, 1990.

IV Reliability Coefficient and Standard Error of Measurement in Classical Mental Test Theory Predicted in the Context of Latent Trait Models

By virtue of the population-free characteristic of the test information function $I(\theta)$, 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 sense of classical mental test theory for each and every specified group of examinees who have taken the same test (cf. Samejima, 1977b, 1987)! This is further facilitated by the proposal of the modifications of the test information function, which use the MLE bias function (cf. Samejima, 1987, 1990), and have been introduced in the preceding chapter.

Thus now we are in the position to 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, *which are tailored for a specific population of examinees.* It will be shown in this chapter how we can do that.

[IV.1] General Case

Let θ_V^* be any estimator of ability θ. We can write

$$\theta_V^* = \theta + \varepsilon \ , \tag{4.1}$$

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

$$\begin{cases} \theta_{V1}^* = \theta + \varepsilon_1 \\ \theta_{V2}^* = \theta + \varepsilon_2 \ , \end{cases} \tag{4.2}$$

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:

$$Cov.(\varepsilon_1, \varepsilon_2) = 0 \ , \tag{4.3}$$

$$Var.(\varepsilon_1) = Var.(\varepsilon_2) \tag{4.4}$$

and

$$Cov.(\theta, \varepsilon_1) = Cov.(\theta, \varepsilon_2) = 0 \ , \tag{4.5}$$

then we will have

$$Corr.(\theta_{V1}^*, \theta_{V2}^*) = [Var.(\theta_{V1}^*) - Var.(\varepsilon_1)][Var.(\theta_{V1}^*)]^{-1} \ . \tag{4.6}$$

Note that if we replace ability θ by the true test score T, a transformed form of θ specific to a given test, and use the observed test score X as the estimator of T, and E as its error of estimation, then (4.1) can be rewritten in the form

(4.7) $$X = T + E \ ,$$

which represents the fundamental assumption in classical mental test theory, and (4.6) becomes a familiar formula for the reliability coefficient $r_{X_1X_2}$,

(4.8) $$r_{X_1X_2} = Var.(T)[Var.(X)]^{-1} \ .$$

In classical mental test theory, however, researchers seldom check if these assumptions are acceptable. In fact, in many cases (4.5) is violated if we replace θ by T , and ε_1 and ε_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

(4.9) $$\begin{aligned} Var.(\varepsilon) &= E[\varepsilon - E(\varepsilon)]^2 \\ &= E[\varepsilon - E(\varepsilon \mid \theta)]^2 + E[E(\varepsilon \mid \theta) - E(\varepsilon)]^2 \\ &\quad + 2E[(\varepsilon - E(\varepsilon \mid \theta))(E(\varepsilon \mid \theta) - E(\varepsilon))] \ . \end{aligned}$$

This indicates that, if the error variable ε is conditionally unbiased for the interval of θ of interest, then (4.9) will be reduced to the form

(4.10) $$Var.(\varepsilon) = E[\varepsilon^2] \ .$$

[IV.2] Reliability Coefficient of a Test in the Sense of Classical Mental Test Theory When the Maximum Likelihood Estimator of θ Is Used

Let $\hat{\theta}_V$ or $\hat{\theta}$ denote the maximum likelihood estimator of θ based upon the response pattern V . If: 1) $\hat{\theta}$ is conditionally unbiased for the interval of θ 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 θ , by the normal distribution $N(\theta, [I(\theta)]^{-1/2})$ for the interval of θ within which the examinees' ability practically distributes. Thus we have from (4.10)

(4.11) $$Var.(\varepsilon) \doteq E[\{I(\theta)\}^{-1}] \ .$$

When this is the case, from (4.6) we can write

(4.12) $$Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - E[\{I(\theta)\}^{-1}]][Var.(\hat{\theta}_1)]^{-1} \ .$$

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.

The appropriateness of the above normal approximation of the conditional distribution of $\hat{\theta}$, given θ, 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 θ 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 (4.5) should be accepted.

If the conditional unbiasedness is *not* supported, i.e., if $B(\hat{\theta}_V \mid \theta)$ does not approximately equal zero for all values of θ in the interval of interest, however, then we shall be able to adopt one of the modified test information functions, $\Upsilon(\theta)$ or $\Xi(\theta)$. Thus we can rewrite (4.12) into the forms

$$Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - E[\{\Upsilon(\theta)\}^{-1}]][Var.(\hat{\theta}_1)]^{-1} \tag{4.13}$$

and

$$Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - E[\{\Xi(\theta)\}^{-1}]][Var.(\hat{\theta}_1)]^{-1} \quad . \tag{4.14}$$

We can decide which of the modified formulae, (4.13) or (4.14), is more appropriate to use in a specified situation.

[IV.3] Standard Error of Measurement of a Test in the Sense of Classical Mental Test Theory When the Maximum Likelihood Estimator of θ Is Used

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. It is usually represented by the reciprocal of the square root of the 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

$$S.E. = E[\{I(\theta)\}^{-1/2}] \tag{4.15}$$

when the conditions 1) and 2) described in the preceding section are met, and by

$$S.E.1 = E[\{\Upsilon(\theta)\}^{-1/2}] \tag{4.16}$$

or

$$S.E.2 = E[\{\Xi(\theta)\}^{-1/2}] \tag{4.17}$$

otherwise.

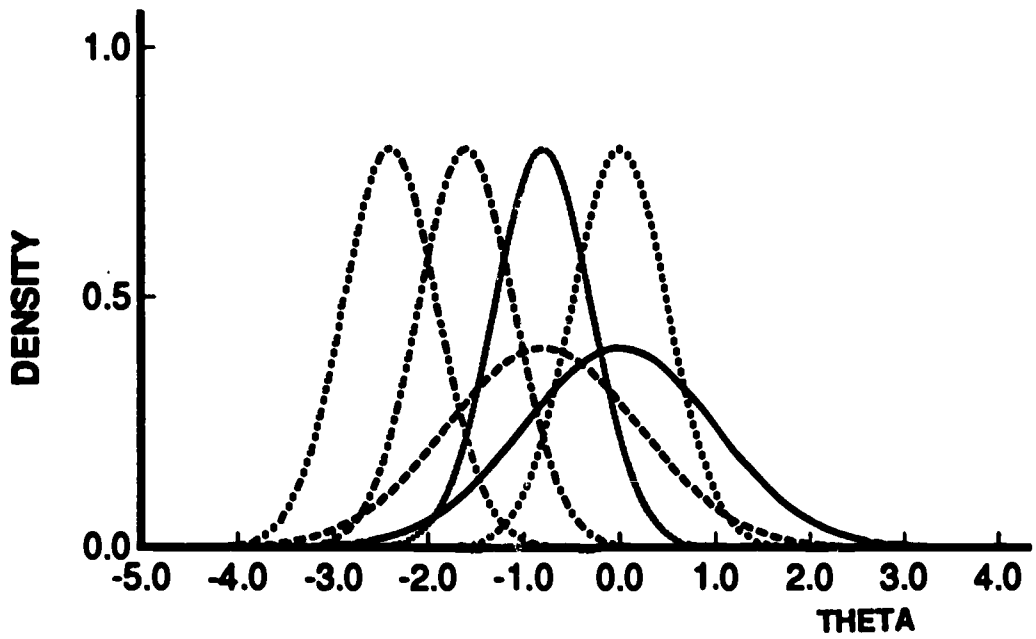

FIGURE 4-1

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).

[IV.4] Examples

For the purpose of illustration, six ability distributions are hypothesised, 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 (4.12), (4.13), (4.14), (4.15), (4.16) and (4.17). 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 4-1 presents the density functions of these six distributions. The hypothetical test used here is the same one introduced in the preceding chapter, which consists of thirty equivalent dichotomous items following the logistic model represented by (3.43) with the common values of parameters, $a_g = 1.0$ and $b_g = 0.0$, respectively, and with the scaling factor D set equal to 1.7. The MLE bias function and the square roots of the test information function $I(\theta)$ and of its two modification formulae $\Upsilon(\theta)$ and $\Xi(\theta)$ of this test are shown in Figure 3-5 of the preceding chapter.

Tables 4-1 and 4-2 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 θ 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 θ 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 (4.12) distributes

TABLE 4-1

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 θ for Each Population Are Also Given.

POPULATION	RELIABILITY 1	RELIABILITY 2	RELIABILITY 3	MEAN OF THETA	VARIANCE OF THETA
1	0.89641	0.78053	0.76629	0.00000	0.98322
2	0.82324	0.26479	0.25256	-0.80000	0.98322
3	0.81738	0.80074	0.79920	0.00000	0.25155
4	0.73250	0.66611	0.65589	-0.80000	0.25155
5	0.47715	0.21681	0.20093	-1.60000	0.25155
6	0.20049	0.01182	0.01109	-2.40000	0.25155

TABLE 4-2

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 θ for Each Population Are Also Given.

POPULATION	STAND.ERROR 1	STAND.ERROR 2	STAND.ERROR 3	MEAN OF THETA	VARIANCE OF THETA
1	0.30548	0.37648	0.38514	0.00000	0.98322
2	0.37887	0.64293	0.66397	-0.80000	0.98322
3	0.23521	0.24717	0.24811	0.00000	0.25155
4	0.29172	0.32802	0.33326	-0.80000	0.25155
5	0.48839	0.73440	0.76583	-1.60000	0.25155
6	0.91974	2.76394	2.88922	-2.40000	0.25155

TABLE 4-3

Three Theoretical Variances of the Maximum Likelihood Estimates of θ 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 θ for Each Population Are Also Given.

POPULATION	VARIANCE OF MLE 1	VARIANCE OF MLE 2	VARIANCE OF MLE 3	MEAN OF THETA	VARIANCE OF THETA
1	1.09684	1.25968	1.28308	0.00000	0.98322
2	1.19432	3.71324	3.89296	-0.80000	0.98322
3	0.30775	0.31414	0.31475	0.00000	0.25155
4	0.34341	0.37763	0.38352	-0.80000	0.25155
5	0.52718	1.16023	1.25189	-1.60000	0.25155
6	1.25469	21.28788	22.68190	-2.40000	0.25155

TABLE 4-4

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 θ for Each Population Are Also Given.

POPULATION	VARIANCE OF ERROR 1	VARIANCE OF ERROR 2	VARIANCE OF ERROR 3	MEAN OF THETA	VARIANCE OF THETA
1	0.11363	0.27646	0.29987	0.00000	0.98322
2	0.21111	2.73003	2.90974	-0.80000	0.98322
3	0.05620	0.06260	0.06320	0.00000	0.25155
4	0.09186	0.12609	0.13197	-0.80000	0.25155
5	0.27563	0.90868	1.00034	-1.60000	0.25155
6	1.00314	21.03633	22.43035	-2.40000	0.25155

TABLE 4-5

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 Covariances Are Also Presented.

POPULATION	RELIABILITY	MEAN 1	MEAN 2	VARIANCE 1	VARIANCE 2	COVARIANCE
1	0.90788	-0.00311	0.00106	1.19069	1.16769	1.07051
2	0.69812	-0.81435	-0.80971	1.07982	1.09703	0.96663
3	0.80724	0.00785	-0.00754	0.33578	0.33443	0.27051
4	0.72334	-0.85777	-0.84349	0.40504	0.39310	0.28863
5	0.55304	-1.68722	-1.67511	0.42299	0.40820	0.22980
6	0.32187	-2.28115	-2.25897	0.21639	0.23189	0.07210

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 θ 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 (4.13) using $\Upsilon(\theta)$ instead of $I(\theta)$ indicates a substantial reduction from the one obtained by (4.12) for each of the six ability distributions. The reduction is especially conspicuous for the populations 2, 5, and 6, whose ability distributes on lower levels of θ where the discrepancies between $I(\theta)$ and $\Upsilon(\theta)$ are large. Among the six populations the predicted reliability coefficient obtained by means of (4.13) varies from 0.012 to 0.781, showing and even larger range than that obtained by (4.12). Similar results were obtained for the predicted reliability coefficient given by (4.14), 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 (4.13) is relatively small, as is expected from the second graph of Figure 3-5.

As for the standard error of measurement, we can see in Table 4-2 that similar results were obtained, only in reversed order, of course. In classical mental test theory, the standard error of measurement σ_E is given by

$$\sigma_E = [Var.(X)]^{1/2}[1 - r_{X_1X_2}]^{1/2} \quad , \tag{4.18}$$

where, as before, $r_{X_1X_2}$ indicates the reliability coefficient. Comparison of Table 4-1 and Table 4-2 reveals that there are substantial discrepancies between the values of σ_E obtained by formula (4.18) using the *tailored* reliability coefficients in Table 4-1, which are based upon the maximum likelihood estimate $\hat{\theta}$, in place of $r_{X_1X_2}$ in (4.18) and the corresponding standard errors of measurement, which were obtained by formulae (4.15) through (4.17) and presented in Table 4-2. To give some examples, for Population No. 1 the results of (4.18) 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 θ and the three theoretical error variances are presented in Tables 4-3 and 4-4, respectively, for each of the six hypothetical populations. The latter were obtained by (4.11) and by replacing $I(\theta)$ in (4.11) by $\Upsilon(\theta)$ and $\Xi(\theta)$, respectively, and the former are the sum of these separate error variances and the variance of θ .

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 θ 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 4-5. Comparison of each of these results with the corresponding three *tailored* reliability coefficients in Table 4-1 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 a 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 θ 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, $\Upsilon(\theta)$. The next most 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 the 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 $\Upsilon(\theta)$, 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 the correlation coefficients in Table 4-5 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.

[IV.5] Discussion and Conclusions

Test information function $I(\theta)$ and its two modification formulae, $\Upsilon(\theta)$ and $\Xi(\theta)$, have been used to predict the reliability coefficient and the standard error of measurement which are *tailored* for each specific ability distribution. Examples of the prediction have been given and a simulation study has been conducted and shown for comparison. These examples using equivalent test items have been rather intentionally chosen to make the differences among the separate ability distributions, and those among the three predicted indices for each ability distribution, clearly visible.

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

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

[2] Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 1977b, 233-247.

[3] Samejima, F. Bias function of the maximum likelihood estimate of ability for discrete item responses. *ONR/RR-87-1*, 1987.

[4] Samejima, F. Modifications of the test information function. *ONR/RR-90-1*, 1990.

V Validity Measures in the Context of Latent Trait Models

From the scientific point of view, we need to confirm if a given test indeed measures what it is supposed to measure, even if we have chosen our items carefully enough in regard to their contents, and even if we are equipped with highly sophisticated mathematics.

By virtue of the population-free nature of latent trait theory, we should be able to find some indices of item validity, and of test validity, which are not affected by the group of examinees. The resulting indices should not be *incidental* as those in classical mental test theory are, but truly be attributes of the item and the test themselves. Thus an attempt has been made in the present research to obtain such population-free measures of item validity and of test validity, which are basically *locally* defined.

[V.1] Performance Function: Regression of the External Criterion Variable on the Latent Variable

It is assumed that there exists an external criterion variable, which can be measured directly or indirectly. This is the situation which is also assumed when we deal with *criterion-oriented validity* or *predictive validity* in classical mental test theory.

Let γ denote the *criterion variable*, representing the performance in a specific job, etc. We shall consider the conditional density of the criterion performance, given ability, and denote it by $\xi(\gamma \mid \theta)$. The *performance function*, $\varsigma(\theta)$, can be defined as the regression of γ on θ, or by taking, say, the 75, 90 or 95 percentile point of each conditional distribution of γ, given θ. Let p_a denote the probability which is large enough to satisfy us as a confidence level. Thus we can write

$$p_a = \int_{\varsigma(\theta)}^{\overline{\gamma}} \xi(\gamma \mid \theta)\, d\gamma \ , \tag{5.1}$$

where $\overline{\gamma}$ denotes the least upper bound of the criterion variable γ.

Figure 5-1 illustrates the relationships among θ, γ, p_a, $\xi(\gamma \mid \theta)$ and $\varsigma(\theta)$. It may be reasonable to assume that the functional relationship between θ and $\varsigma(\theta)$ is relatively simple, not as is illustrated by the solid line in Figure 5-2, i.e., we do not expect $\varsigma(\theta)$ to go up and down frequently within a relatively short range of θ. We shall assume that $\varsigma(\theta)$ is twice differentiable with respect to θ.

In dealing with an additional dimension or dimensions in latent space, i.e., the criterion variable or variables, one of the most difficult issues is to keep the population-free nature, which is characteristic of the latent trait models, the main feature that distinguishes the theory from classical mental test theory, among others. If we consider the projection of the operating characteristic of a discrete item response on the criterion dimension, for example, then the resulting operating characteristic as a function of γ has to be *incidental*, for it has to be affected by the population distribution of θ.

We need to start from the conditional distribution of γ, given θ, therefore, which can be conceived of as being intrinsic in the relationship between the two variables, and independent of the population distribution of θ. We assume that $\varsigma(\theta)$ takes on the same value only at a finite or an enumerable number of points of . . Let $P^*_{k_g}(\varsigma)$ be the conditional probability assigned to the discrete response k_g, given ς. We can write

$$P^*_{k_g}(\varsigma) = \sum_{\varsigma(\theta)=\varsigma} P_{k_g}(\theta) \quad . \tag{5.2}$$

[V.2] When $\varsigma(\theta)$ Is Strictly Increasing in θ : Simplest Case

The simplest case is that $\varsigma(\theta)$ is strictly increasing in θ. In this case, $\varsigma(\theta)$ has a one-to-one correspondence with θ, and (5.2) becomes simplified into the form

$$P^*_{k_g}(\varsigma) = P^*_{k_g}[\varsigma(\theta)] = P_{k_g}(\theta) \quad . \tag{5.3}$$

If, in addition, $\{\partial\theta/\partial\varsigma\}$ is finite throughout the entire range of θ, then we obtain

$$\frac{\partial}{\partial\varsigma} P^*_{k_g}(\varsigma) = [\frac{\partial}{\partial\theta} P_{k_g}(\theta)] \frac{\partial\theta}{\partial\varsigma} \quad . \tag{5.4}$$

Let $I^*_{k_g}(\varsigma)$ be the item response information function defined as a function of ς. We can write

$$\begin{aligned} I^*_{k_g}(\varsigma) &= -\frac{\partial^2}{\partial\varsigma^2} \log P^*_{k_g}(\varsigma) = -\frac{\partial}{\partial\varsigma}[\{\frac{\partial}{\partial\theta} \log P_{k_g}(\theta)\} \frac{\partial\theta}{\partial\varsigma}] \\ &= I_{k_g}(\theta) \, (\frac{\partial\theta}{\partial\varsigma})^2 - [\frac{\partial}{\partial\theta} P_{k_g}(\theta)] \, [P_{k_g}(\theta)]^{-1} \, \frac{\partial^2\theta}{\partial\varsigma^2} \quad . \end{aligned} \tag{5.5}$$

Let $I^*_g(\varsigma)$ and $I^*(\varsigma)$ be the amounts of information given by a single item g and by the total test, respectively, for a fixed value of ς. Then we have from (2.3), (2.8) and (5.5)

$$I^*_g(\varsigma) = E[I^*_{k_g}(\varsigma) \mid \varsigma] = \sum_{k_g} I^*_{k_g}(\varsigma) \, P^*_{k_g}(\varsigma) = I_g(\theta) \, (\frac{\partial\theta}{\partial\varsigma})^2 \tag{5.6}$$

and

$$I^*(\varsigma) = \sum_{g=1}^{n} I^*_g(\varsigma) = I(\theta) \, (\frac{\partial\theta}{\partial\varsigma})^2 \quad . \tag{5.7}$$

If we take the square roots of these two information functions defined for ς, then we obtain

$$[I^*_g(\varsigma)]^{1/2} = [I_g(\theta)]^{1/2} \frac{\partial\theta}{\partial\varsigma} \tag{5.8}$$

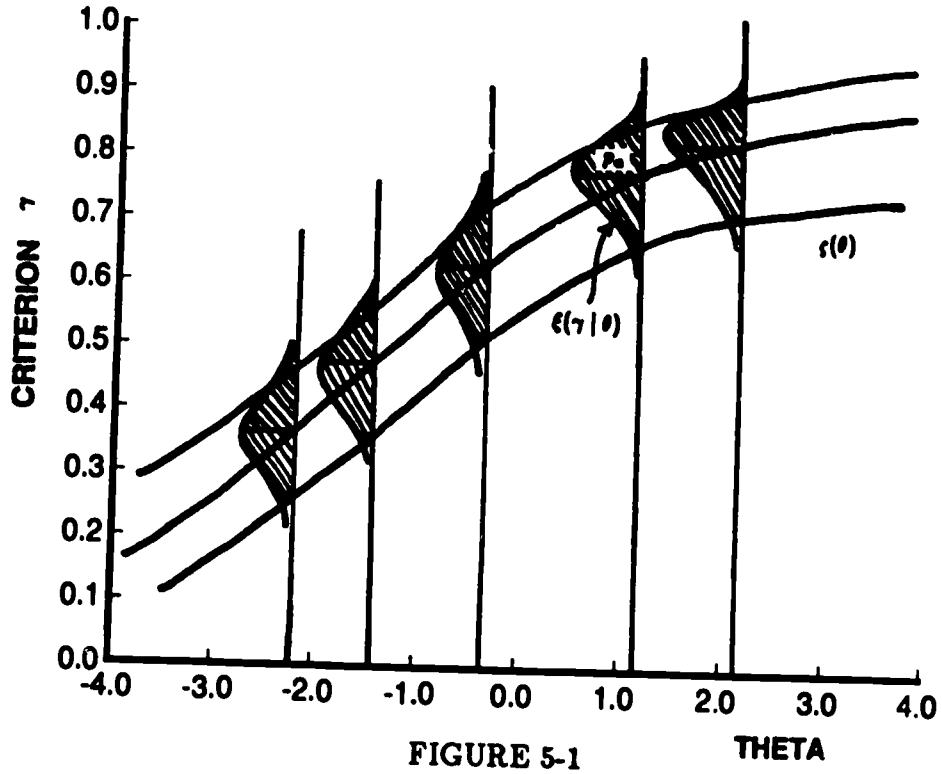

FIGURE 5-1

Relationships among θ, γ, p_a, $\xi(\gamma \mid \theta)$ and $\zeta(\theta)$.

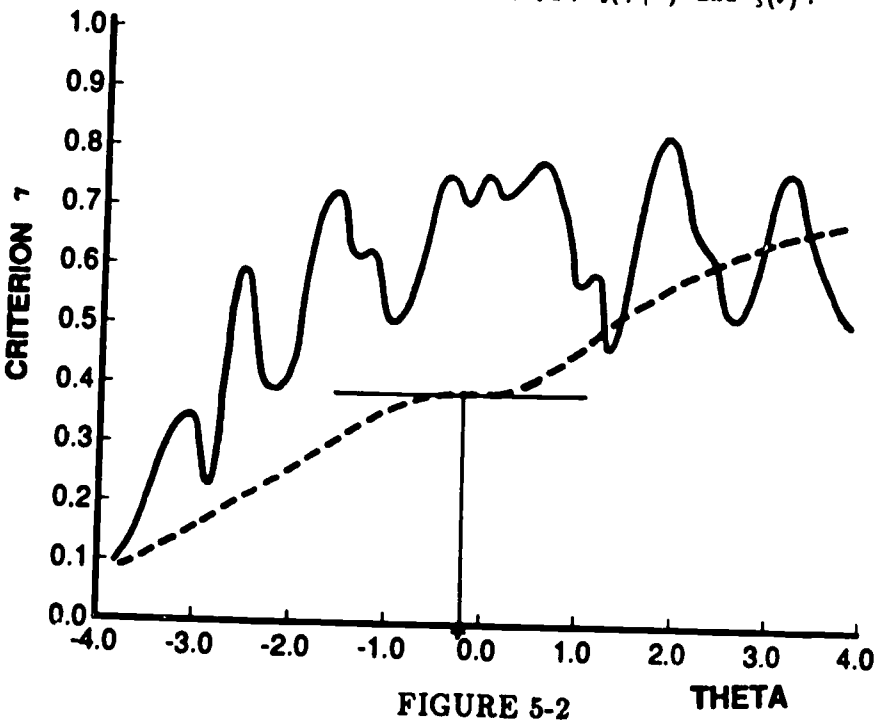

FIGURE 5-2

Two Hypothetical Performance Functions $\zeta(\theta)$, One of Which Is Not Likely to Be the Case (Solid Line), and the Other Has a Derivative Equal to Zero at One Point of θ (Dashed Line).

and

$$[I^*(\varsigma)]^{1/2} = [I(\theta)]^{1/2} \frac{\partial \theta}{\partial \varsigma} \quad . \tag{5.9}$$

Since a certain constant nature exists for the square root of the item information function while the same is not true with the original item information function (cf. Samejima, 1979, 1982), $[I_g^*(\varsigma)]^{1/2}$ given by (5.8) instead of the original function given by (5.6) may be more useful in some occasions. This will be discussed later in this section, when the validity in selection *plus* classification is discussed.

Suppose that we have a critical value, γ_0, of the criterion variable, which is needed for succeeding in a specified job, and that we try to accept applicants whose values of the criterion variable are γ_0 or greater. If our primary purpose of testing is to make an accurate selection of applicants, then (5.8) and (5.9) for $\varsigma = \gamma_0$, or their squared values shown by (5.6) and (5.7), indicate item and test validities, respectively. If for some item formula (5.8) or (5.6) assumes a high value at $\varsigma = \gamma_0$, then the standard error of estimation of ς around $\varsigma = \gamma_0$ becomes small and chances are slim that we make misclassifications of the applicants by accepting unqualified persons and rejecting qualified ones, and the reversed relationship holds when (5.8) or (5.6) assumes a low value at $\varsigma = \gamma_0$. The same logic applies to the total test by using formula (5.9) or (5.7) instead of (5.8) or (5.6).

It should be noted in (5.8) or in (5.9), that $[I_g^*(\gamma_0)]^{1/2}$ or $[I^*(\gamma_0)]^{1/2}$ consists of two factors, i.e., 1) the square root of the item information function $I_g(\theta)$ or that of the test information function $I(\theta)$ and 2) the partial derivative of ability θ with respect to ς at $\varsigma = \gamma_0$. These two factors in each formula are *independent* of each other, i.e., one belongs to the item or to the test and the other to the statistical relationship between θ and γ. We also notice that these two factors are in a *supplementary* relationship. Thus while it is important to have a large amount of item information, or of test information, it is even more so to have large values of the derivative, $\{\partial\theta/\partial\varsigma\}$, in the vicinity of $\varsigma = \gamma_0$, for this will increase the amount of item information defined with respect to ς uniformly in that vicinity, and also that of test information, as is obvious from the right hand sides of (5.8) and (5.9). In other words, it is desirable for the purpose of selection *for ς to increase slowly in θ in the vicinity of $\varsigma = \gamma_0$*.

Since, in general, the same ability θ has predictabilities for more than one kind of job performance, or of potential of achievement, *the performance function varies for different criterion variables*. Note that neither $[I_g(\theta)]^{1/2}$ nor $[I(\theta)]^{1/2}$ is changed even when the criterion variable is switched. Thus, for a fixed item or test whose amount of information is reasonably large around $\varsigma = \gamma_0$, the derivative $\{\partial\theta/\partial\varsigma\}$ in the vicinity of $\varsigma = \gamma_0$ determines the appropriateness of the use of the item or of the test for the purpose of selection with respect to a specific job, etc. If this derivative assumes a high value, then an item or a test which provides us with a medium amount of information may be acceptable for our purpose of selection, while we will need an item or a test whose amount of information is substantially larger if the derivative is low. Also for the same criterion variable γ the derivative $\{\partial\theta/\partial\varsigma\}$ varies for different values of γ_0, so the appropriateness of an item or of a test depends upon our choice of γ_0, too. The above logic also applies for the formulae (5.6) and (5.7), i.e., for the case in which we choose the information functions, instead of their square roots, changing $\{\partial\theta/\partial\varsigma\}$ to its squared value.

It is obvious from (5.6) and (5.8) that we can choose either $I_g(\theta(\gamma_0))$ or $[I_g(\theta(\gamma_0))]^{1/2}$ for use in item selection, for their rank orders across different items are identical, and they equal the rank orders of $I_g^*(\gamma_0)$ as well as those of $[I_g^*(\gamma_0)]^{1/2}$.

If we take another standpoint that our purpose of testing is *not only to make a right selection of applicants but also to predict the degree of success in the job for each selected individual*, then we will need to integrate $[I_g^*(\varsigma)]^{1/2}$ and $[I^*(\varsigma)]^{1/2}$, respectively, since we must estimate ς accurately not only around $\varsigma = \gamma_0$ but also for $\varsigma > \gamma_0$. If we choose $[I_g^*(\varsigma)]^{1/2}$ and $[I^*(\varsigma)]^{1/2}$ in preference to their squared values, we will obtain from (5.8) and (5.9)

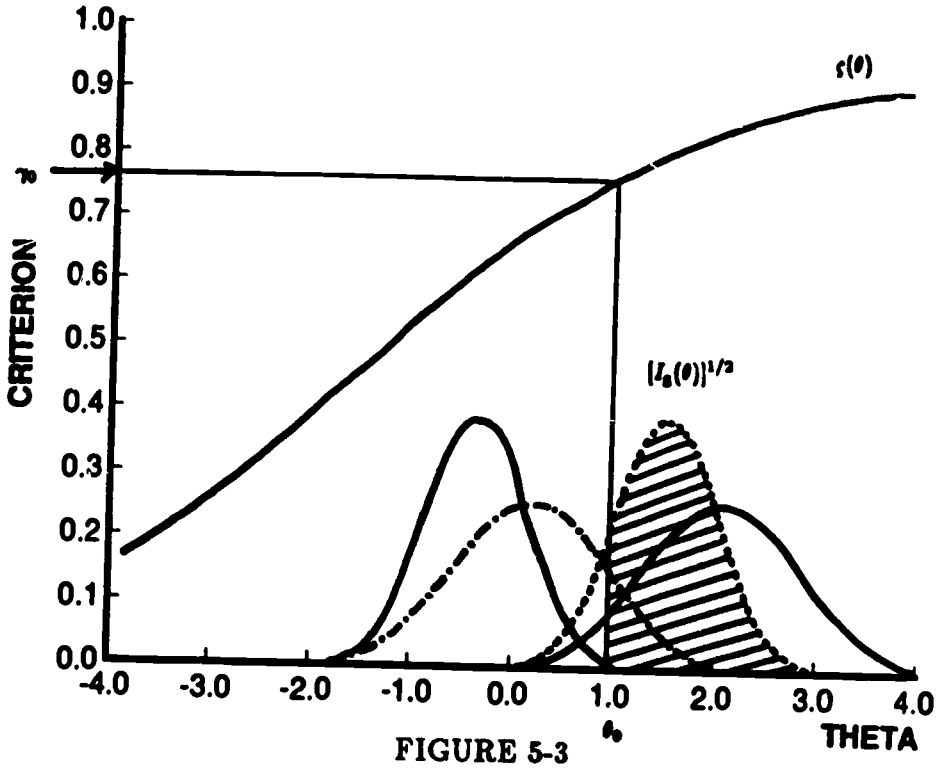

FIGURE 5-3

Some Examples of the Relationship between γ_0 and the Item Validity Measure Given by (5.10).

(5.10) $$\int_{\Omega_\varsigma} [I_g^*(\varsigma)]^{1/2}\, d\varsigma = \int_{\Omega_\theta} [I_g(\theta)]^{1/2}\, d\theta$$

and

(5.11) $$\int_{\Omega_\varsigma} [I^*(\varsigma)]^{1/2}\, d\varsigma = \int_{\Omega_\theta} [I(\theta)]^{1/2}\, d\theta \ ,$$

where Ω_ς and Ω_θ indicate the domains of ς and θ for which $\varsigma(\theta) \geq \gamma_0$, respectively. In this situation we need to select items which assume high values of (5.10) instead of (5.8), or a test which provides us with a high value of (5.11) in place of (5.9). Note that formulae (5.10) and (5.11) imply that we can obtain these two validity measures directly from the original item and test information functions, respectively, i.e., *without actually transforming* θ *to* ς, as long as we can identify the domain Ω_θ. This is true for any criterion variable γ.

Some examples illustrating the values of (5.10) are given in Figure 5-3 for hypothetical items. In the simplest case observed in this section and illustrated in Figures 5-1 and 5-3, these two domains, Ω_θ and Ω_ς, are provided by the two intervals, $(\theta_0\ ,\ \infty)$ and $(\gamma_0\ ,\ \overline{\gamma})$, where

(5.12) $$\theta_0 = \theta(\gamma_0)$$

and $\overline{\gamma}$ denotes the least upper bound of γ.

It should be noted that the above pair of validity measures depends upon our choice of the critical value γ_0. If this value is low, i.e., a specified job does not require high levels of competence with

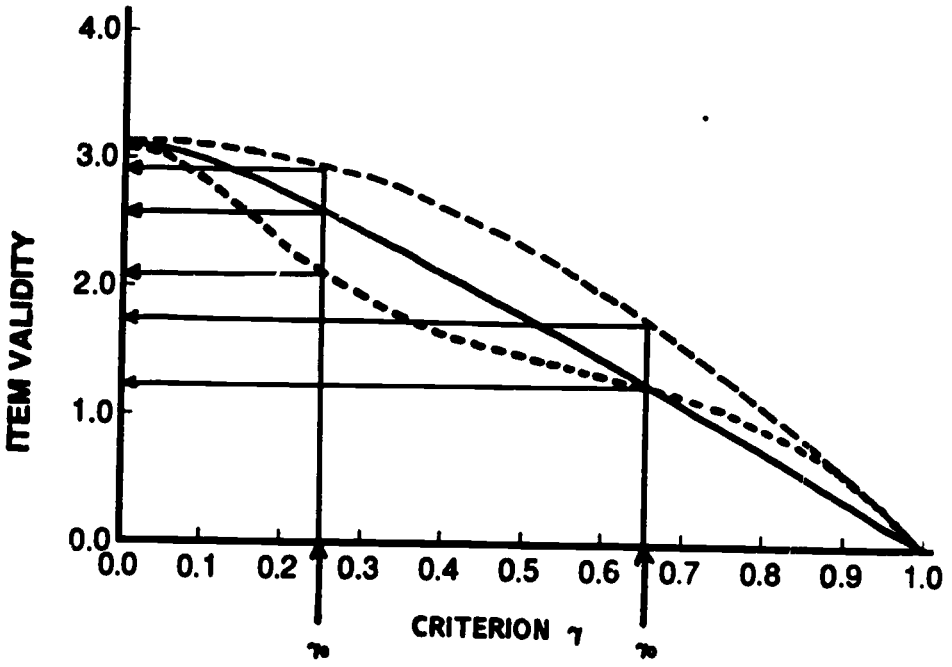

FIGURE 5-4

Relationship between γ_0 and Item Validity Indicated by (5.10) for Three Hypothetical Dichotomous Items Whose Operating Characteristics for the Correct Answer Are Strictly Increasing with Zero and Unity as Their Asymptotes.

respect to the criterion variable γ, then these validity indices assume high values, and vice versa. It has been pointed out (Samejima, 1979, 1982) that there is a certain constancy in the amount of information provided by a single test item. To give an example, if an item is dichotomously scored and has a strictly increasing operating characteristic for success with zero and unity as its two asymptotes, then the area under the curve for $[I_g(\theta)]^{1/2}$ equals π, *regardless of the mathematical form of the operating characteristic and its parameter values.* We can see, therefore, that if our items belong to this type then the functional relationship between γ_0 and the item validity measure given by (5.10) will be *monotone decreasing*, with π and zero as its two asymptotes, for each and every item. Figure 5-4 illustrates this relationship for three hypothetical items of this type. As we can see in this figure, the appropriateness of the items changes with γ_0 in an absolute sense, and also relatively to other items with γ_0, and the rank orders of desirability among the items depend upon our choice of γ_0.

We can see from (5.10) that this validity measure necessarily assumes a high value if an item is difficult, and the same applies to (5.11) for the total test. This implies that these validity measures alone cannot indicate the desirability of an item and of a test precisely *for a specific population of examinees.* In selecting items or a test, therefore, it is desirable to take the ability distribution of the examinees into account, if the information concerning the ability distribution of a *target* population is more or less available. In so doing we shall be able to avoid choosing items which are too difficult for the target population of examinees. Let $f(\theta)$ denote the density function of the ability distribution for a specific population of examinees, and $f^*(\zeta)$ be that of ζ for the same population. Then we can write

$$f^*(\zeta) = f(\theta)\frac{\partial\theta}{\partial\zeta} \quad . \tag{5.13}$$

Adopting this as the weight function, from (5.8) and (5.9) we obtain as the validity indices *tailored* for a specific population of examinees

(5.14) $$\int_{\Omega_\varsigma} [I_g^*(\varsigma)]^{1/2} f^*(\varsigma)\, d\varsigma = \int_{\Omega_\theta} [I_g(\theta)]^{1/2} f(\theta) \frac{\partial \theta}{\partial \varsigma}\, d\theta$$

and

(5.15) $$\int_{\Omega_\varsigma} [I^*(\varsigma)]^{1/2} f^*(\varsigma)\, d\varsigma = \int_{\Omega_\theta} [I(\theta)]^{1/2} f(\theta) \frac{\partial \theta}{\partial \varsigma}\, d\theta \ .$$

Thus by using (5.14) and (5.15) instead of (5.10) and (5.11) we shall be able to make appropriate item selection and test selection *for a target population* or sample, provided that the information concerning its ability distribution is more or less available. Note that, unlike (5.10) and (5.11), formulae (5.14) and (5.15) imply that these validity measures are also heavily dependent upon the functional formula of $\varsigma(\theta)$.

If we choose to use the area under the curve of the information function instead of that of its square root, we obtain from (5.6) and (5.7)

(5.16) $$\int_{\Omega_\varsigma} I_g^*(\varsigma)\, d\varsigma = \int_{\Omega_\theta} I_g(\theta) \frac{\partial \theta}{\partial \varsigma}\, d\theta$$

and

(5.17) $$\int_{\Omega_\varsigma} I^*(\varsigma)\, d\varsigma = \int_{\Omega_\theta} I(\theta) \frac{\partial \theta}{\partial \varsigma}\, d\theta \ ,$$

respectively. We notice that in this case, unlike those of (5.10) and (5.11), the integrands of the right hand sides of (5.16) and (5.17) are no longer independent of the functional formula of $\varsigma(\theta)$. Also when information about the ability distribution of a target population of examinees is more or less available, the *tailored* item and test validity indices become

(5.18) $$\int_{\Omega_\varsigma} I_g^*(\varsigma)\, f^*(\varsigma)\, d\varsigma = \int_{\Omega_\theta} I_g(\theta)\, f(\theta)\, (\frac{\partial \theta}{\partial \varsigma})^2\, d\theta$$

and

(5.19) $$\int_{\Omega_\varsigma} I^*(\varsigma)\, f^*(\varsigma)\, d\varsigma = \int_{\Omega_\theta} I(\theta)\, f(\theta)\, (\frac{\partial \theta}{\partial \varsigma})^2\, d\theta \ ,$$

respectively, if we choose to use the information functions instead of their square roots.

Note that, unlike the validity measures for *selection* purposes, in the present situation the rank orders of validity across different items, or different tests, depend upon the choice of the validity index. Thus a question is: which of the formulae, (5.10) or (5.16), and (5.11) or (5.17), are better as the item and the test validity indices for *selection plus classification* purposes? A similar question is also addressed with respect to (5.14) and (5.18), and to (5.15) and (5.19). These are tough questions to answer. While the choice of the square root of the item information function has an advantage of a certain constancy which has been observed earlier in this subsection, the use of the item information has a benefit of additivity, i.e., by virtue of (2.8) the sum total of (5.16) over all the item g's equals (5.17), and the same relationship holds between (5.18) and (5.19). The answers to these questions are yet to be searched.

When our purpose of testing is strictly the *classification* of individuals, as in assigning those people to different training programs, in guidance, et ., (5.10) and (5.11), or (5.16) and (5.17), also serve as the

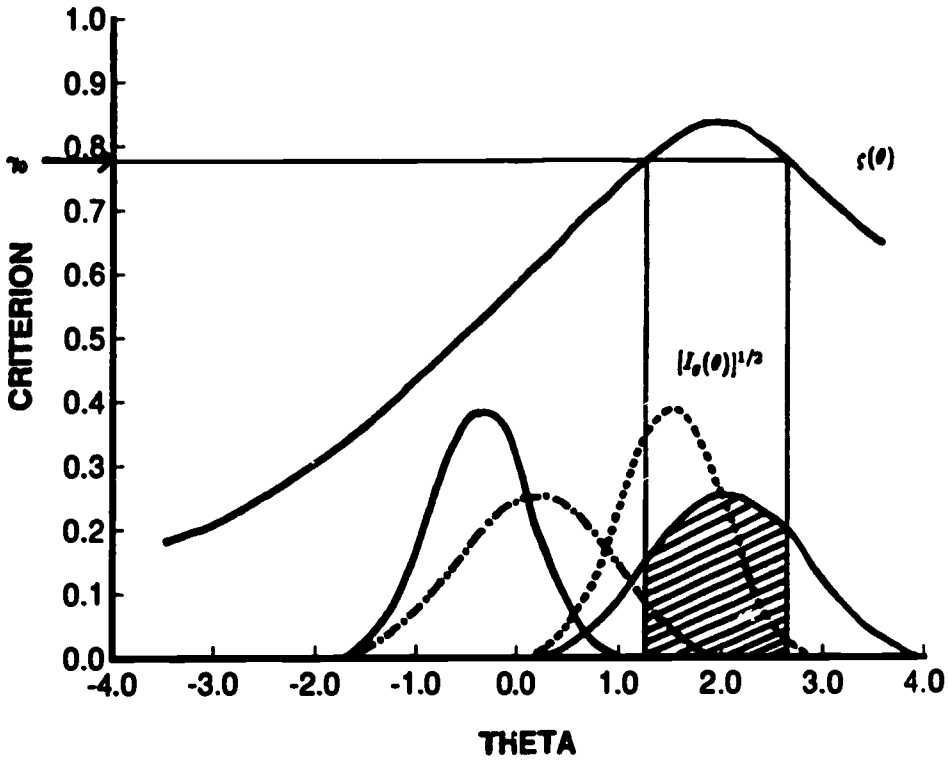

FIGURE 5-5

Example of the Performance Function $\varsigma(\theta)$ Which Is Piecewise Monotone in θ.

validity measures of an item and of a test, respectively. In this case, we must set $\gamma_0 = \underline{\gamma}$ in defining the domains, Ω_ς and Ω_θ, where $\underline{\gamma}$ is the greatest lower bound of γ. Thus the two domains, Ω_ς and Ω_θ, in these formulae become those of ς and θ for which $\underline{\gamma} \leq \varsigma(\theta) \leq \overline{\gamma}$. It is obvious that these formulae provide us with the item and the test validity measures, respectively, for the same reason explained earlier. The same logic applies for the *tailored* validity measures provided by (5.14) and (5.15), and by (5.18) and (5.19), when the information concerning the ability distribution of a target population is more or less available

[V.3] Test Validity Measures Obtained from More Accurate Minimum Variance Bounds

When $\{\partial\varsigma/\partial\theta\} = 0$ at some value of θ, as is illustrated by a dashed line in Figure 5-2, $\{\partial\theta/\partial\varsigma\}$ becomes positive infinity, and so does the item validity measure given by (5.8). This fact provides us with some doubt, for, while we can see that at such a point of ς item validity is high, we must wonder if positive infinity is an adequate measure. It is also obvious from (2.8) that the same will happen to the total test if it includes at least one such item. Our question is: *should we search for more meaningful functions than the item and test information functions?* This topic will be discussed in this section.

Necessity of the search for a more accurate measure than the test information function becomes more urgent when the performance function, $\varsigma(\theta)$, is not strictly increasing in θ, but is, say, only piecewise monotone in θ with finite $\{\partial\theta/\partial\varsigma\}$ and differentiable with respect to θ, as is illustrated in Figure 5-5. The illustrated performance function is still simple enough, but indicates the trend that after a certain point of ability the performance level in a specified job decreases. This can happen when the job does not provide enough challenge for persons of very high ability levels.

Since $I^*(\varsigma)$ serves as the reciprocal of the conditional variance of the maximum likelihood estimate of ς only asymptotically and there exist more accurate minimum variance bounds for any (asymptotically) unbiased estimator (cf. Kendall and Stuart, 1961), we can search for more accurate test validity measures than the one given by (5.9) by using the reciprocal of the square roots of such minimum variance bounds.

Let $J_{rs}(\theta)$ be defined as

$$J_{rs}(\theta) = E[\frac{L_V^{(r)}(\theta)}{L_V(\theta)} \frac{L_V^{(s)}(\theta)}{L_V(\theta)} \mid \theta] \qquad r, s = 1, 2, ..., k \tag{5.20}$$

where

$$L_V^{(r)}(\theta) = \frac{\partial^r}{\partial \theta^r} L_V(\theta) = \frac{\partial^r}{\partial \theta^r} P_V(\theta) \ . \tag{5.21}$$

Let $J(\theta)$ denote the $(k \times k)$ matrix of the element $J_{rs}(\theta)$, and $J_{rs}^{-1}(\theta)$ be the corresponding element of its inverse matrix, $J^{-1}(\theta)$. Note that when $k = 1$ we can rewrite (5.20) into the form

$$\begin{aligned} J_{kk}(\theta) &= J_{11}(\theta) = E[\{\frac{\partial}{\partial \theta} \log L_V(\theta)\}^2 \mid \theta] \\ &= -E[\frac{\partial^2}{\partial \theta^2} \log P_V(\theta) \mid \theta] \ , \end{aligned} \tag{5.22}$$

and from this, (2.7) and (2.8) we can see that $J(\theta)$ is a (1×1) matrix whose element is the test information function, $I(\theta)$, itself. A set of improved minimum variance bounds is given by

$$\sum_{r=1}^{k} \sum_{s=1}^{k} \varsigma^{(s)}(\theta) \ J_{rs}^{-1}(\theta) \ \varsigma^{(r)}(\theta) \tag{5.23}$$

(cf. Kendall and Stuart, 1961), where $\varsigma^{(s)}(\theta)$ denotes the s-th partial derivative of $\varsigma(\theta)$ with respect to θ. We obtain, therefore, for a set of new test validity measures

$$[\sum_{r=1}^{k} \sum_{s=1}^{k} \gamma_0^{(s)} \ J_{rs}^{-1}(\theta(\gamma_0)) \ \gamma_0^{(r)}]^{-1/2} \ , \tag{5.24}$$

where $\gamma_0^{(s)}$ indicates the s-th partial derivative of ς with respect to θ at $\varsigma = \gamma_0$.

The use of this new test validity measure will ameliorate the problems caused by $\{\partial \varsigma / \partial \theta\} = 0$, if we choose an appropriate k. The resulting algorithm will become much more complicated, however, and we must expect a substantially larger amount of CPU time for computing these measures when k is greater than unity. Note that (5.24) equals (5.9) when $k = 1$.

[V.4] Multidimensional Latent Space

When our latent space is multidimensional, a generalization of the idea given in Section 5.3 for the unidimensional latent space can be made straightforwardly. We can write

$$\theta = \{ \theta_u \}' \qquad u = 1, 2, ..., \eta \ , \tag{5.25}$$

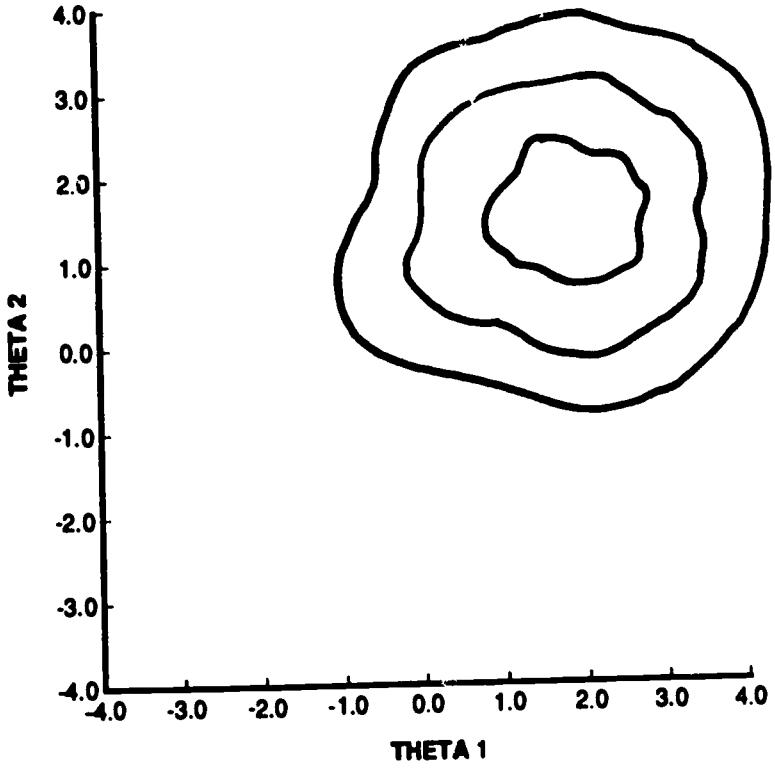

FIGURE 5-6

Area Ω_θ for Different γ_0's in Two-Dimensional Latent Space for a Hypothesized Test.

and the performance function $\varsigma(\theta)$ becomes a function of η independent variables. A minimum variance bound is given by

$$\sum_{u=1}^{\eta}\sum_{v=1}^{\eta} \frac{\partial \varsigma(\theta)}{\partial \theta_u} \frac{\partial \varsigma(\theta)}{\partial \theta_v} I_{uv}^{-1}(\theta) \ , \tag{5.26}$$

where $I_{uv}^{-1}(\theta)$ is the (u,v)-th element of the inverse matrix of the $(\eta \times \eta)$ symmetric matrix, whose element is given by

$$I_{uv}(\theta) = E[\frac{1}{L} \frac{\partial L}{\partial \theta_u} \frac{\partial L}{\partial \theta_v} \mid \theta] \tag{3.27}$$

with L abbreviating $L_V(\theta)$, or $P_V(\theta)$. The reciprocal of the square root of (5.27) will provide us with the counterpart of (5.9) for the multidimensional latent space. For $\eta = 2$, the area Ω_θ may look like one of the contours illustrated in Figure 5-6, depending upon our choice of γ_0, taking the axis for γ vertical to the plane defined by θ_1 and θ_2.

In a more complex situation where both ability and the criterion variables are multidimensional, we must consider the projection of the item information function on the criterion subspace from the ability subspace, in order to have the item validity function for each item, and then the test validity function. It is anticipated that we must deal with a higher mathematical complexity in such a case. The situation

will substantially be simplified, however, if the total set of items consists of several subsets of items, each of which measures, exclusively, a single ability dimension and a single criterion dimension.

[V.5] Discussion and Conclusions

Some considerations have been made concerning the validity of a test and that of a single item. Effort has been focused upon searching for measures which are population-free, and which will provide us with local and abundant information just as the information functions do in comparison with the test reliability coefficient in classical mental test theory. In so doing, validity indices for different purposes of testing and also those which are tailored for a specific population of examinees have been considered.

The above considerations for the item and test validities may be just part of many possible approaches. We may still have a long way to go before we discover the most useful measures of the item and test validities. The present research may stimulate other researchers so that they will pursue this topic further, taking different approaches.

We notice that the test validity measures proposed in this research can be modified by using one of the two modification formulae, $\Upsilon(\theta)$ and $\Xi(\theta)$, of the test information function (cf. Chapter 3), in place of the original $I(\theta)$. This will be investigated in the future, when the characteristics of these two modification formulae have further been investigated and clarified.

References

[1] Kendall, M. G. and Stuart, A. *The advanced theory of statistics. Vol. 2.* New York: Hafner, 1961.

[2] Samejima, F. Constant information model: A new promising item characteristic function. *ONR/RR-79-1*, 1979.

[3] Samejima, F. Information loss caused by noise in models for dichotomous items. *ONR/RR-82-1*, 1982.

VI Further Investigation of the Nonparametric Approach to the Estimation of the Operating Characteristics of Discrete Item Responses

In the present research a method has been proposed which increases accuracies of estimation of the operating characteristics of discrete item responses, while pertaining to the two features described in Section 2.3, and the new procedure has been tested upon dichotomous items. It has proved to be effective, especially when the true operating characteristic is represented by a *steep* curve, and also at the lower and upper ends of the ability distribution where the estimation tends to be inaccurate because of smaller numbers of subjects involved in the base data. Tentatively, it is called the *Differential Weight Procedure*, and it belongs to the Conditional P.D.F. Approach (cf. Chapter 2). This procedure costs more CPU time than the Simple Sum Procedure, which has been used frequently (cf. Samejima, 1981, 1988), but the advantage of handling more than one item, say, fifty, together in the Conditional P.D.F. Approach is still there.

[VI.1] Simple Sum Procedure of the Conditional P.D.F. Approach Combined with the Normal Approach Method

It is obvious from the discussion given in Chapter 2 that the Conditional P.D.F. Approach combined with the Normal Approach Method is the simplest and one of the most economical procedures in CPU time. Out of the three procedures of the Conditional P.D.F. Approach the Simple Sum Procedure is the simplest one (cf. Samejima, 1981). For this reason, the combination of the Simple Sum Procedure of the Conditional P.D.F. Approach and the Normal Approach Method has most frequently been applied for simulated and empirical data. Fortunately, in spite of the simplicity of the procedure, the results with simulated data in the adaptive testing situation and with simulated and empirical data in the paper-and-pencil testing situation indicate that we can estimate the operating characteristics fairly accurately by using this combination (cf. Samejima, 1981, 1984). This seems to prove the *robustness* of the Conditional P.D.F. Approach. For one thing, there is a good reason why Normal Approach Method works well, for the conditional distribution of τ, given $\hat{\tau}$, is indeed normal if the (unconditional) distribution of τ is normal, and it is a truncated normal distribution if the (unconditional) distribution of τ is rectangular, and the truncation is negligible for most of the conditional distributions.

In the Simple Sum Procedure of the Conditional P.D.F. Approach, the operating characteristic, $P_{k_g}(\theta)$, of the discrete item response k_g of an *unknown* item g is estimated through the formula

$$\hat{P}_{k_g}(\theta) = \hat{P}^*_{k_g}[\tau(\theta)] = \sum_{s \in k_g} \phi(\tau \mid \hat{\tau}_s)[\sum_{s=1}^{N} \phi(\tau \mid \hat{\tau}_s)]^{-1} \quad , \tag{6.1}$$

where s $(= 1, 2, \ldots, N)$ indicates an individual examinee, and $\phi(\tau \mid \hat{\tau}_s)$ denotes the conditional density of τ, given $\hat{\tau}_s$. This conditional density is estimated by using the estimated conditional moments of τ, given $\hat{\tau}_s$, using one of the four methods, as was described in Section 2.3.

In the Weighted Sum Procedure of the Conditional P.D.F. Approach, we have for the estimated operating characteristic of k_g

$$\hat{P}_{k_g}(\theta) = \hat{P}^*_{k_g}[\tau(\theta)] = \sum_{s \in k_g} w(\hat{\tau}_s)\phi(\tau \mid \hat{\tau}_s)[\sum_{s=1}^{N} w(\hat{\tau}_s)\phi(\tau \mid \hat{\tau}_s)]^{-1} \tag{6.2}$$

where $w(\hat{\tau}_s)$ is the weight function of $\hat{\tau}_s$. When we combine one of these two approaches with the Normal Approach Method, $\phi(\tau \mid \hat{\tau}_s)$ in (6.1) or in (6.2) is approximated by the normal density function, using the first two estimated conditional moments of τ, given $\hat{\tau}_s$, which are given by (2.13) and (2.14), respectively, as its parameters, $\mu_{\hat{\tau}_s}$ and $\sigma_{\hat{\tau}_s}$, in the formula

$$\phi(\tau \mid \hat{\tau}_s) = [2\pi]^{-1/2}[\sigma_{\hat{\tau}_s}]^{-1} \exp[-(\tau - \mu_{\hat{\tau}_s})^2/\{2\sigma^2_{\hat{\tau}_s}\}] \quad . \tag{6.3}$$

[VI.2] Differential Weight Procedure

If we accept the approximation of the conditional distribution of $\hat{\tau}$, given τ, by the asymptotic normality, as we do in these approaches (cf. Samejima, 1981), the other conditional distribution, i.e., that of τ, given $\hat{\tau}$, will become more or less *incidental*. Thus in the Bivariate P.D.F. Approach the bivariate distribution of τ and $\hat{\tau}$ is approximated for each separate item score subpopulation of subjects of each unknown test item. In the Conditional P.D.F. Approach, however, the incidentality of this second conditional distribution is not rigorously considered, and the implicit assumption exists such that for the fixed value of $\hat{\tau}$ the conditional distributions of τ are similar for the different item score subpopulations.

Take the dichotomous response level, for example. On this level, each item is scored "right" or "wrong", "affirmative" or "negative", etc. The above assumption of non-incidentality may be acceptable when the operating characteristic of the correct answer of the item is represented by a *mildly steep* curve, as is the case with most practical situations, and the questions are asked to subjects whose ability levels are *compatible* with the difficulty levels of the questions, as is the case with adaptive testing and, though less rigorously, with many cases of paper-and-pencil testing.

This assumption is not acceptable, however, when the operating characteristic of the correct answer is represented by a *steep* curve. If the operating characteristic follows the Guttman scale, for example, then the conditional distributions of τ, given θ, for the two separate item score subpopulations are distinctly separated, and they do not even overlap! If we use the Simple Sum Procedure or the Weighted Sum Procedure for an item which nearly follows the Guttman scale, therefore, the resulting estimated operating characteristics of the correct and the incorrect answers will tend to be *flatter* than they actually are.

This problem can be solved by estimating *differential* conditional distributions of τ, given $\hat{\theta}$, for the separate discrete item responses to an "unknown" item. Let $\phi_{k_g}(\tau \mid \hat{\theta})$ denote the conditional density of τ, given $\hat{\theta}$, for the subpopulation of subjects who share the same discrete item response k_g to an "unknown" item g. We can write

(6.4)
$$\phi_{k_g}(\tau \mid \hat{\theta}) = f^*_{k_g}(\tau)\ \psi(\hat{\theta} \mid \tau)\ [g^*_{k_g}(\hat{\theta})]^{-1} \quad ,$$

where $f^*_{k_g}(\tau)$ indicates the density of τ for the subpopulation of subjects who share k_g as their common item score of item g, $\psi(\hat{\theta} \mid \tau)$ is the conditional density of $\hat{\theta}$, given τ, which is approximated by the normal density, $n[\tau,\ C_1^{-1}]$, and $g^*_{k_g}(\hat{\theta})$ is the marginal density of $\hat{\theta}$, for this subpopulation, and for which we have

(6.5)
$$g^*_{k_g}(\hat{\theta}) = \int_{-\infty}^{\infty} f^*_{k_g}(\tau)\ \psi(\hat{\theta} \mid \tau)\ d\tau \quad .$$

We notice that there is a relationship

(6.6)
$$f^*_{k_g}(\tau) = f^*(\tau)\ P^*_{k_g}(\tau)\ [\int_{-\infty}^{\infty} f^*(\tau)\ P^*_{k_g}(\tau)\ d\tau]^{-1} \quad ,$$

where $f^*(\tau)$ denotes the density of τ for the total population. Since we have

(6.7)
$$\phi(\tau \mid \hat{\theta}) = f^*(\tau)\ \psi(\hat{\theta} \mid \tau)\ [g^*(\hat{\theta})]^{-1} \quad ,$$

where $g^*(\hat{\theta})$ is the density of $\hat{\theta}$ for the total population of subjects which is given by

(6.8)
$$g^*(\hat{\theta}) = \int_{-\infty}^{\infty} f^*(\tau)\ \psi(\hat{\theta} \mid \tau)\ d\tau \quad ,$$

from the above formulae we obtain

(6.9)
$$\phi_{k_g}(\tau \mid \hat{\theta}) = \phi(\tau \mid \hat{\theta})\ P^*_{k_g}(\tau)\ h(\hat{\theta}) \quad ,$$

where $h(\hat{\theta})$ is a function of $\hat{\theta}$ and constant for a fixed value of $\hat{\theta}$. Thus $\phi_{k_g}(\tau \mid \hat{\theta})$ is a density function proportional to $\phi(\tau \mid \hat{\theta})\, P^*_{k_g}(\tau)$. We notice that $\phi(\tau \mid \hat{\theta})$ in this formula is common to all the item scores and across different unknown items, while $P^*_{k_g}(\tau)$ is a specific function of τ for each k_g. Since $\phi(\tau \mid \hat{\theta})$ can be estimated by one of the four methods described in Section 2.3, our effort should be focused on finding an appropriate differential weight function for each k_g. Let $W_{k_g}(\tau)$ denote such a *differential weight function*, which replaces $P^*_{k_g}(\tau)\, h(\hat{\theta})$ in (6.9). Thus we can revise (6.1) and (6.2) into the forms

$$\hat{P}_{k_g}(\theta) = \hat{P}^*_{k_g}[\tau(\theta)] = \sum_{s \in k_g} W_{k_g}(\tau)\phi(\tau \mid \hat{\theta}_s)[\sum_{s=1}^{N} W_{k_g}(\tau; s)\phi(\tau \mid \hat{\theta}_s)]^{-1} \tag{6.10}$$

and

$$\hat{P}_{k_g}(\theta) = \hat{P}^*_{k_g}[\tau(\theta)] = \sum_{s \in k_g} w(\hat{\theta}_s)W_{k_g}(\tau)\phi(\tau \mid \hat{\theta}_s)[\sum_{s=1}^{N} w(\hat{\theta}_s)W_{k_g}(\tau; s)\phi(\tau \mid \hat{\theta}_s)]^{-1} \quad . \tag{6.11}$$

Since the differential weight function $W_{k_g}(\tau)$ involves $P^*_{k_g}(\tau)$, which itself is the target of estimation, we may use its estimate, $\hat{P}^*_{k_g}(\tau)$, obtained by the Simple Sum Procedure or by the Weighted Sum Procedure, as its substitute. In so doing, we may need some local smoothings of $\hat{P}^*_{k_g}(\tau)$ where the estimation involves substantial amounts of error because of locally small numbers of subjects in the base data, etc. In some cases we may need several iterations by renewing the differential weight functions on each stage until the resulting estimated operating characteristic converges.

[VI.3] Examples

We have tried this proposed method on the simulated data provided by Dr. Charles Davis of the Office of Naval Research, using the Simple Sum Procedure of the Conditional P.D.F. Approach combined with the Normal Approach Method with some modifications as the initial estimate of $P_{k_g}(\tau)$ in the differential weight function. These data are simulated on-line item calibration data of the initial itempool calibration based upon conventional testing, in which 100 dichotomous items are divided into four subtests of 25 items each, and each subtest has been administered to 6,000 hypothetical examinees, and those of different rounds based upon adaptive testing, in which each of the 50 new binary items has been administered to a subgroup of 1,500 hypothetical subjects out of the total of 15,000. These hypothetical examinees' ability distributes unimodally within the interval of θ, (-3.0, 3.0), with slight negative skewness.

For the purpose of illustration, Figure 6-1 presents the results of the Differential Weight Procedure using the results of the Simple Sum Procedure of the Conditional P.D.F. Approach combined with the Normal Approach Method with some modifications as the initial estimates, for a couple of items of the initial itempool. They are dichotomous items, and were intentionally selected from those items whose true operating characteristics of the correct answer are non-monotonic, in order to visualize the benefit of the nonparametric estimation of the operating characteristic. In each graph, also presented for comparison is the best fitted operating characteristic of the correct answer following the three-parameter logistic model, which has been given by Dr. Michael Levine. We can see in these graphs that the resulting estimated operating characteristics are fairly close to the true ones, and that they reflect the non-monotonicities. The reader is directed to ONR/RR-90-4 (Samejima, 1990) for more examples.

[VI.4] Sensitivities to Irregularities of Weight Functions

As we have proceeded, several factors have been identified and observed which affect the resulting estimated operating characteristics substantially. They are concerned with the differential weight func-

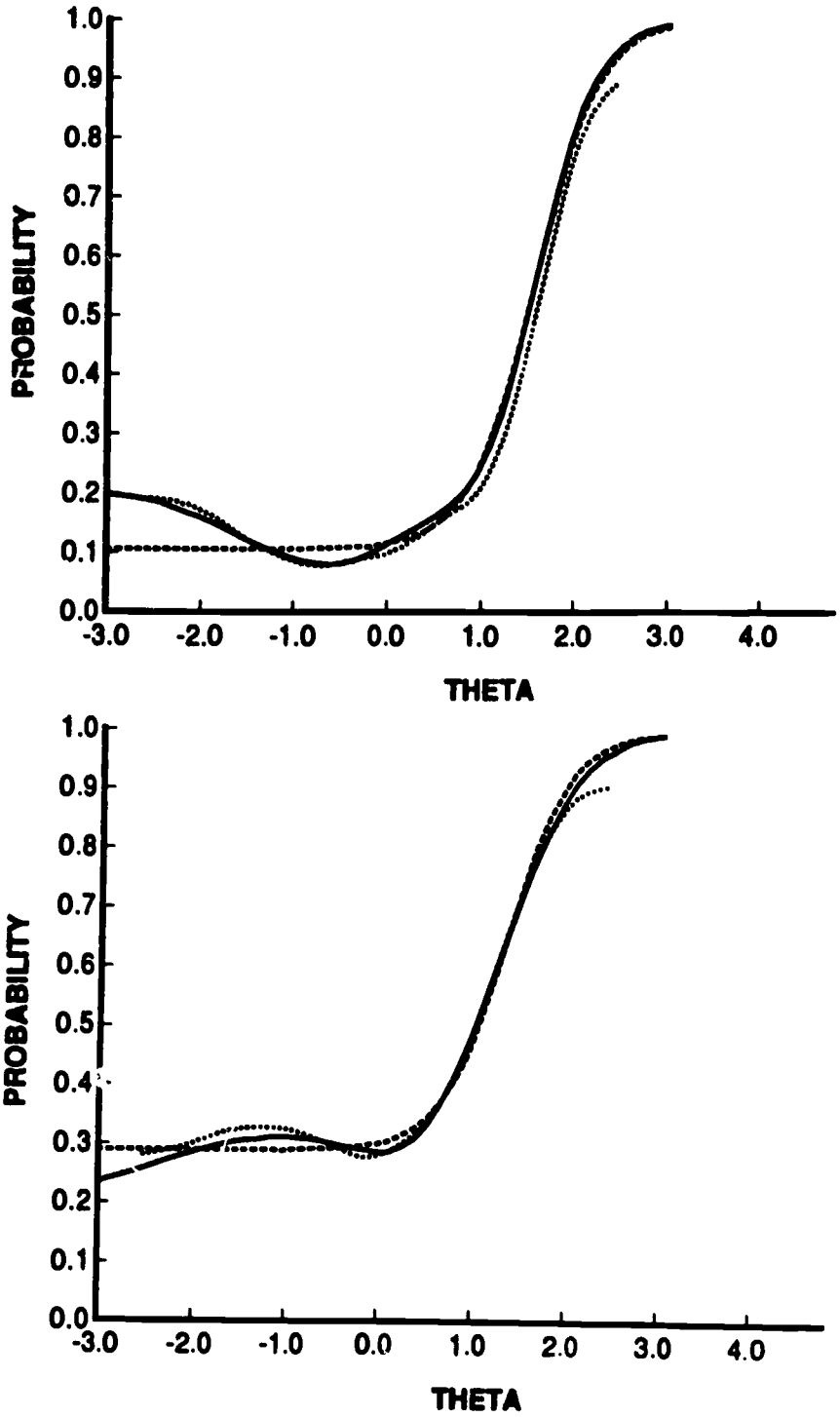

FIGURE 6-1

Two Examples of the Estimated Operating Characteristic of the Correct Answer Using the Differential Weight Procedure (Dotted Line), in Comparison with the True Operating Characteristic (Solid Line) and the Best Fitted Three-Parameter Logistic Curve (Dashed Line).

tion, and can be itemised as: 1) lower end ambiguities, 2) upper end ambiguities, 3) local irregularities and 4) overall irregularities.

Out of these factors, lower and upper end ambiguities basically come from the fact that we do not usually have sufficiently large numbers of subjects on the lowest and the highest ends of the interval of θ of interest upon which the estimation of the operating characteristics is made. Also the fact that the test information function $I(\theta)$ is used in the transformation of θ to τ which is specified by (2.9) may have something to do with these ambiguities. It has been observed (Samejima, 1979b) that in using equivalent items following the Constant Information Model (Samejima, 1979a) the speed of convergence of the conditional distribution of the maximum likelihood estimate $\hat{\theta}$, given θ, to the asymptotic normality with θ and $[I(\theta)]^{-1/2}$ as its two parameters substantially differs for different levels of θ, *in spite of the fact that the amount of test information is constant for every level of* θ. To be more specific, the convergence is observed to be much slower at those levels which are close to either end of the interval of θ for which the amount of test information is non-zero and constant, and faster at intermediate levels of θ. This situation can be ameliorated if we replace the test information function $I(\theta)$ in (2.9) by one of its two modified forms (cf. Chapter 3), $\Upsilon(\theta)$ and $\Xi(\theta)$.

By irregularity we mean non-smoothness, which is exemplified by an unnatural angle, etc. It has been observed that for most items the resulting operating characteristic is amazingly sensitive to these irregularities of the differential weight function. In order to observe these sensitivities, Figure 6-2 illustrates how these irregularities, which are involved in the differential weight function, affect the resulting estimated operating characteristic. For more examples, the reader is directed to ONR/RR-90-4 (Samejima, 1990).

The effect of *local* irregularities is most interesting to observe in the three examples presented by Figure 6-2. In each of these graphs, the artificially *irregular* differential weight function for the correct answer is drawn by a short dashed line, and, in order to emphasise its irregularities, it was proportionally enlarged and shown by a long dashed line. We can see in each graph that, when the differential weight function has an unnatural angle, for example, the resulting estimated operating characteristic of the correct answer also shows an unnatural angle at approximately the same level of θ. We can also see in these graphs how *overall* irregularities of the differential weight function affect the resulting estimated operating characteristic, and how sensitive the latter is to the former. This type of sensitivity of the resulting estimated operating characteristic to the irregularities of the differential weight function is encouraging as well as threatening, for it promises success in the estimation provided that we succeed in finding the right differential weight function.

During the present research period, perhaps the author and her research assistants have spent the greatest amount of time for developing this method, Differential Weight Procedure of the Conditional P.D.F. Approach. Thus, in addition to the results exemplified in this section and in ONR/RR-90-4 (Samejima, 1990), there have been produced so many other results, using different strategies in specifying differential weight functions, etc. The research will be continued in the future, and those results which are not introduced in this final report will be included in the basis upon which the future research will be founded and planned, and will eventually be introduced in future research reports.

[VI.5] Discussion and Conclusions

A new procedure of nonparametric estimation of the operating characteristics of discrete item responses has been proposed, which is called Differential Weight Procedure of the Conditional P.D.F. Approach. Some examples have been given, and sensitivities of the resulting estimated operating characteristics to irregularities of the differential weight functions have been observed and discussed. These outcomes suggest the importance of further investigation of the weight function in the future.

To summarise, although Simple Sum Procedure of the Conditional P.D.F. Approach combined with

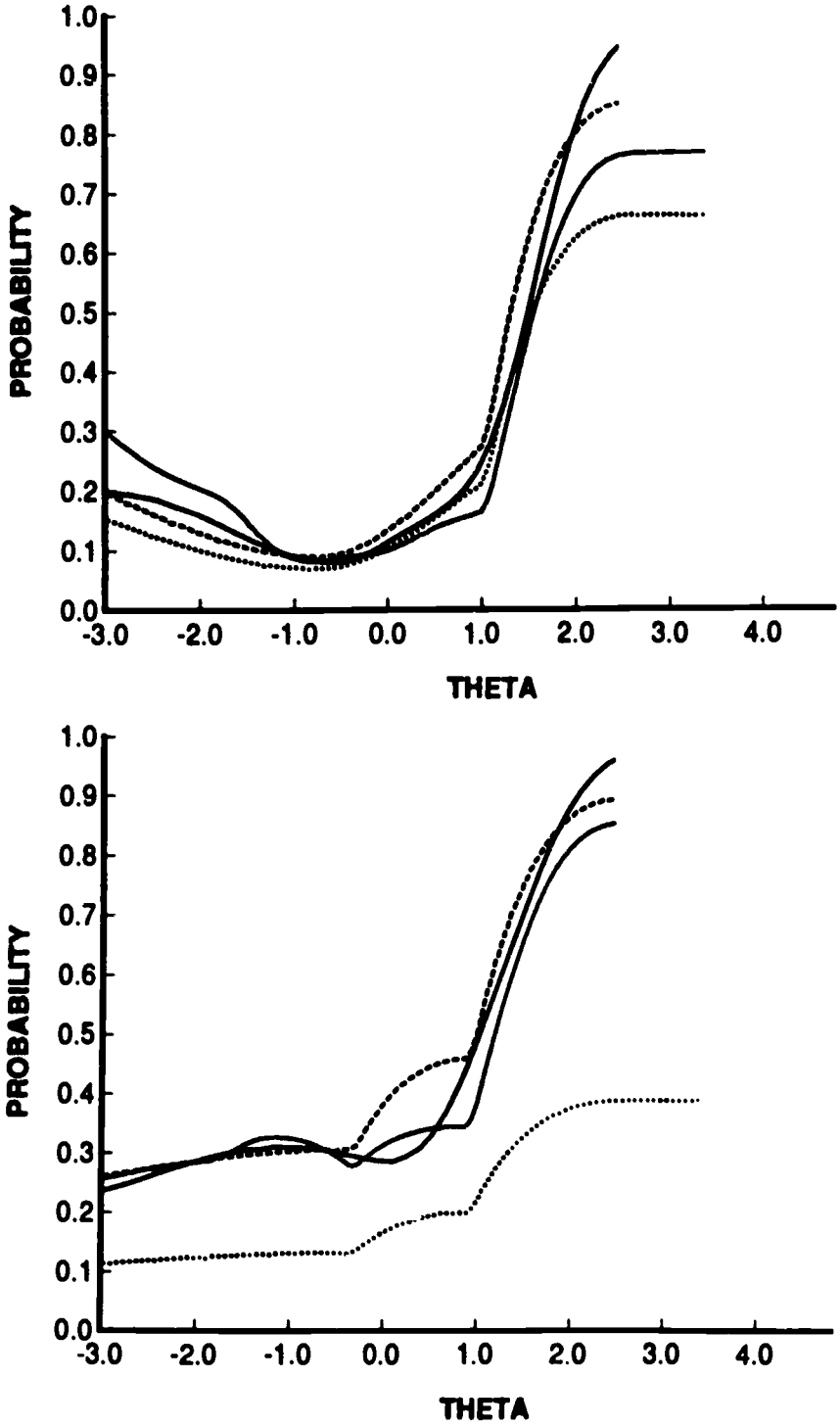

FIGURE 6-2

Three Examples of the Estimated Operating Characteristic of the Correct Answer Using the Differential Weight Procedure (Dotted Line), in Comparison with the True Operating Characteristic (Solid Line), When the Differential Weight Function (Short Dashed Line) Has Irregularities. The Function Was Also Proportionally Enlarged and Plotted (Long Dashed Line) to Visualise the Angles and Other Irregularities Well.

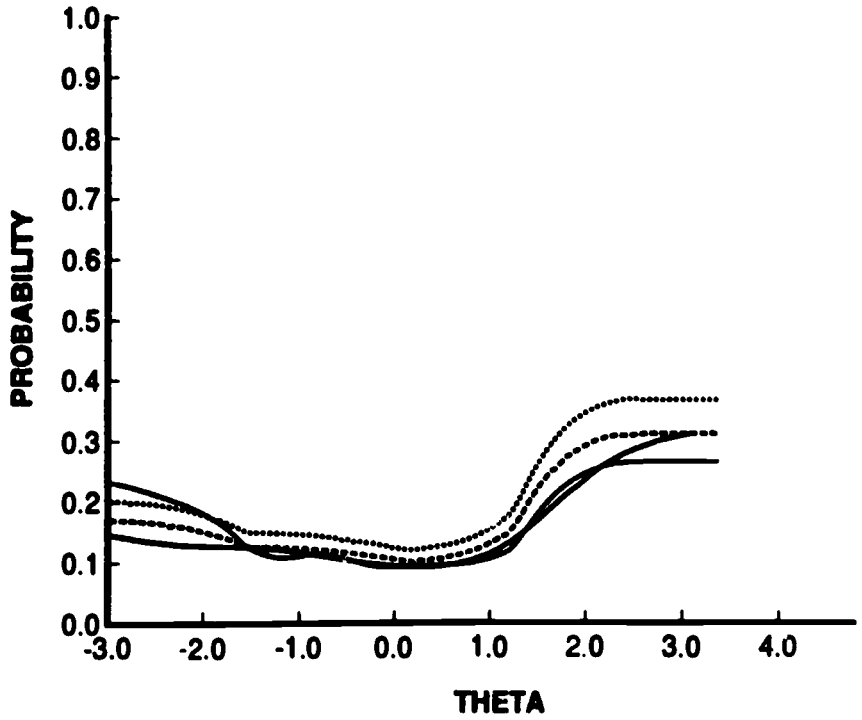

FIGURE 6-2 (Continued)

the Normal Approach Method works reasonably well for the on-line item calibration of adaptive testing, and also for the paper-and-pencil testing, especially when the number of subjects is large, if we wish to increase the accuracy of estimation we can use the Differential Weight Procedure. The disadvantage will be the added CPU time, so we need to consider the balance of the cost and accuracy of estimation before we make our decision. It will be less expensive, however, if we compare the CPU time required for the present procedure with the time required for the Bivariate P.D.F. Approach.

References

[1] Samejima, F. Constant information model: A new, promising item characteristic function. *ONR/RR-79-1*, 1979a.

[2] Samejima, F. Convergence of the conditional distribution of the maximum likelihood estimate, given latent trait, to the asymptotic normality: Observations made through the constant information model. *ONR/RR-79-3*, 1979b.

[3] Samejima, F. Final Report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item. *Final Report of N00014-77-C-0360*, Office of Naval Research, 1981.

[4] Samejima, F. Plausibility functions of Iowa Vocabulary Test items estimated by the Simple Sum Procedure of the Conditional P.D.F. Approach. *ONR/RR-84-1*, 1984.

[5] Samejima, F. Final Report: Advancement of latent trait theory. *Final Report of N00014-81-C-0569*, Office of Naval Research, 1988.

[6] Samejima, F. Differential Weight Procedure of the Conditional P.D.F. Approach for estimating the operating characteristics of discrete item responses. *ONR/RR-90-4*, 1990.

VII Content-Based Observation of Informative Distractors and Efficiency of Ability Estimation

Partly because of the availability of computer software, such as Logist (Wingersky, Barton and Lord, 1982), Bilog (Bock and Atkin, 1981), etc., it is a common procedure among researchers that they *mold* the operating characteristics of correct answers into the three-parameter logistic model, ignoring their possible non-monotonicity. In some cases, strategies are even taken so that distractors, which cause the non-monotonicity, are considered as undesirable and are replaced by some other *non-threatening* alternative answers.

A question must be raised as to whether this strategy is wise. In this chapter, this issue will be discussed both from theory and from practice, and a new strategy of writing test items, which leads to more efficient ability estimation, will be proposed. It will take advantage of the ease in handling mathematics attributed to parameterization, and yet minimize the effect of noise caused by random guessing.

[VII.1] Non-Monotonicity of the Conditional Probability of the Positive Response, Given Latent Variable

This section deals basically with the essence or a summary of the paper published by the author more than twenty years ago (Samejima, 1968), as one of the research reports of the L. L. Thurstone Psychometric Laboratory of the University of North Carolina. The content of the paper was a protocol which led to the proposal of a *new* family of models for the multiple-choice test item (Samejima, 1979b). The author believes that this paper published in 1968 still gives *new ideas* to today's research communities.

The paper is concerned with the nominal response, and also multiple-choice situations, in which examinees are required to choose one of the given alternatives. in connection with the graded response model (cf. Samejima, 1969, 1972). For a multiple-choice item a certain number of false answers are given in addition to the correct answer. In a general case it is impossible to score them in a graded manner in accordance with their degrees of attainment toward the goal. Thus the multiple-choice situation should be treated as a special instance of the nominal level of response, although, in addition, the problem of random or irrational choice should be investigated.

Confining discussions to examinees who have responded to item g incorrectly, there can be diversity of false answers if they have responded to it freely, without being forced to choose one of a set of alternative answers. It is conceivable that some of the false answers may require high levels of ability measured while some others may not, some may be related to the ability measured strongly while some others may not, etc. An objective measure of the plausibility of a specified false answer is its operating characteristic, i.e., the probability of its occurrence defined for a fixed value of ability θ, and, therefore, expressed as a function of θ.

Let $M_s(\theta)$ be a sequence of the conditional probabilities corresponding to the cognitive subprocesses required in finding the plausibility of response k_g to item g, and $U_{k_g}(\theta)$ be the conditional probability that an examinee discovers the irrationality of response k_g as the answer to item g, on condition that he has already found out its plausibility. The operating characteristic of k_g, which is denoted by $P_{k_g}(\theta)$, can be expressed by

$$P_{k_g}(\theta) = [1 - U_{k_g}(\theta)] \prod_{s \in k_g} M_s(\theta) \quad , \tag{7.1}$$

since it is reasonably assumed that an examinee who gives a response k_g to item g is one who has succeeded in finding k_g's plausibility, and yet failed in finding its irrationality. We notice that this

formula is exactly the same in its structure as the definition of $P_{z_g}(\theta)$ on the graded response level, where $M_s(\theta)$ is replaced by $M_i(\theta)$ and $U_{k_g}(\theta)$ is replaced by $M_{(z_g+1)}(\theta)$ (cf. Samejima, 1972). Defining $M_{k_g}(\theta)$ such that

$$M_{k_g}(\theta) = \prod_{s \in k_g} M_s(\theta) \ , \tag{7.2}$$

we can rewrite (7.1) into

$$P_{k_g}(\theta) = M_{k_g}(\theta)[1 - U_{k_g}(\theta)] \ . \tag{7.3}$$

It will reasonably be assumed from their definitions that both $M_{k_g}(\theta)$ and $U_{k_g}(\theta)$ be strictly increasing in θ, provided that a specified response k_g is a *good mistake* in the sense that the discoveries of its plausibility and irrationality are properly related with ability θ. It will also be reasonably assumed that the upper asymptotes of $M_{k_g}(\theta)$ and $U_{k_g}(\theta)$ are unity, and the lower asymptote of $M_{k_g}(\theta)$ is zero.

We assume that both $M_{k_g}(\theta)$ and $U_{k_g}(\theta)$ are three-times-differentiable with respect to θ. It is easily observed that, in order to satisfy the unique maximum condition (Samejima, 1969, 1972), $P_{k_g}(\theta)$ defined by (7.3) must fulfill the following inequalities:

$$\frac{\partial^2}{\partial\theta^2} \log M_{k_g}(\theta) = \frac{\partial}{\partial\theta}[\frac{\partial}{\partial\theta} M_{k_g}(\theta)\{M_{k_g}(\theta)\}^{-1}] < 0 \tag{7.4}$$

and

$$\frac{\partial^2}{\partial\theta^2} \log[1 - U_{k_g}(\theta)] = \frac{\partial}{\partial\theta}[-\frac{\partial}{\partial\theta} U_{k_g}(\theta)\{1 - U_{k_g}(\theta)\}^{-1}] < 0 \ . \tag{7.5}$$

(For proof, see Samejima, 1968.) Note that in this case the lower asymptote of $U_{k_g}(\theta)$ need not be zero. The operating characteristic of a specified response k_g which satisfies the unique maximum condition was called the *plausibility curve* (Samejima, 1968), and later the *plausibility function* (cf. Samejima, 1984a). As the condition suggests, the plausibiiity curve is necessarily unimodal. A schematised hypothesis for the plausibility curve is the following. The probability that an examinee will find the plausibility, but will *fail* in discovering the irrationality, of a specified response k_g as the answer to item g is a function of ability θ; it increases as ability θ increases, reaches maximum at a certain value of θ, and then decreases afterwards. If an item provides many such responses, their plausibility curves will be powerful sources of information in estimating examinees' abilities. That is to say, we can make use of specific wrong answers to an item as sources of information, as well as the correct answer.

Let $P_g(\theta)$ denote the operating characteristic of the correct answer of a dichotomous item g in the free-response situation. Let $P_g^*(\theta)$ be the same function, but in the multiple-choice situation. The conventional three-parameter model is represented by

$$P_g^*(\theta) = c_g + (1 - c_g)P_g(\theta) \ , \tag{7.6}$$

where c_g is the probability with which an examinee will guess correctly (Lord and Novick, 1968). This is a monotonically increasing function of θ with c_g (≥ 0) and unity as its lower and upper asymptotes, provided that $P_g(\theta)$ is strictly increasing in θ with zero and unity as its lower and upper asymptotes.

The psychological hypothesis which has led to the formula (7.6) in the multiple-choice situation is the following. If an examinee has ability θ, then the probability that he will know the correct answer is given by $P_g(\theta)$; if he does not know it, he will guess randomly, and, with probability c_g, will guess correctly (Lord and Novick, 1968). Thus we have for the operating characteristic of the correct answer of item g in the multiple-choice situation

$$P_g(\theta) + [1 - P_g(\theta)]c_g \quad , \tag{7.7}$$

which leads to (7.6). This hypothesis may not necessarily be appropriate for ability measurement. One can never tell in the measurement of a reasoning ability, for instance, whether an examinee *knows* the correct answer to item g or not, until he has tried to solve it. He may respond with an incorrect alternative without guessing at all. To explain such a case we need some other hypothesis than the one which leads to the formula (7.6).

Hereafter, we assume that $P_g(\theta)$ is *strictly increasing* in θ with zero and unity as its lower and upper asymptotes, and is twice-differentiable with respect to θ. Suppose, further, that both $P_g(\theta)$ and $[1 - P_g(\theta)]$ satisfy the unique maximum condition. In this case $P_g^*(\theta)$ defined by (7.6) *does not* satisfy either of Conditions (i) and (ii) for the unique maximum, unless c_g is zero, i.e., the free-response situation, although they are fulfilled for the negative answer to item g (cf. Samejima, 1973). Observations and discussion are made (Samejima, 1968) giving two simple cases of the multiple-choice situation as examples. In those examples, only two items are involved, and the response pattern, (1,0), is solely treated, and precise mathematical derivations are given.

A possible correction for the conventional functional formula for the operating characteristic of the correct answer of a multiple-choice item can be made by introducing the probability of random guessing defined for a fixed value of θ. Let $d_g(\theta)$ denote this probability. A reasonable assumption for this function may be that it be *non-increasing* in θ. Thus the probability with which an examinee of ability θ will answer item g correctly by following the due cognitive process is expressed by $[1 - d_g(\theta)]P_g(\theta)$; and the one with which he will give the correct answer by guessing should be $d_g(\theta)c_g$. For economy of notation, let $P_g^*(\theta)$ be the operating characteristic of the correct answer to item g in the corrected functional formula also. We can write

$$\begin{aligned} P_g^*(\theta) &= [1 - d_g(\theta)]P_g(\theta) + d_g(\theta)c_g \\ &= P_g(\theta) + d_g(\theta)[c_g - P_g(\theta)] \quad . \end{aligned} \tag{7.8}$$

A schematized psychological hypothesis which leads to this formula is as follows. If an examinee has ability θ, then he will depend upon random guessing in answering item g with probability $d_g(\theta)$; in that case, the conditional probability with which he will guess correctly is given by c_g. If he does not depend upon random guessing, he will try to solve the item by the due cognitive process, and will succeed in solving it with probability $P_g(\theta)$. Thus according to this functional formula the probability with which an examinee will respond with an incorrect alternative without guessing is given by $[1 - d_g(\theta)][1 - P_g(\theta)]$, which is nil in the model represented by the formula (7.6).

We can conceive of several factors which may affect the functional formula for $d_g(\theta)$. The difficulty of item g may be one of them; the discriminating power may be another; the number of alternatives attached to item g may also affect the probability, i.e., it may be that the fewer the number of alternatives, the more tempted to depend upon random guessing an examinee will be; also the plausibilities of the alternatives may be counted as a factor.

In a simplified case where $d_g(\theta)$ is constant throughout the whole range of θ, we can rewrite (7.8) in the following form.

(7.9) $$P_g^*(\theta) = d_g c_g + [1 - d_g] P_g(\theta) \quad .$$

This is somewhat similar to formula (7.6), the conventional functional formula for the operating characteristic of the correct answer of a multiple-choice item. The lower asymptote of the present function is $d_g c_g \ (\leq c_g)$, however, while it is c_c in (7.6); the upper asymptote of the present function is $[1 - d_g(1 - c_g)]$, which can be less than unity, while it is unity in (7.6). In a special case where $d_g = 0$, that is, an examinee tries to solve item g by proper reasoning with probability one, (7.9) reduces to $P_g(\theta)$, the operating characteristic of the correct answer in the free-response situation. In another special case where $d_g = 1$, that is, an examinee depends upon random guessing with probability one, (7.9) reduces to a constant, c_g . In the more general case where $d_g(\theta)$ varies as θ varies, it is observed from (7.8) that

(7.10)
$$\begin{cases} 0 < P_g(\theta) \leq P_g^*(\theta) \leq c_g & ; \text{if} \quad \theta < \theta_0 \\ P_g^*(\theta) = c_g = P_g(\theta) & ; \text{if} \quad \theta = \theta_0 \\ c_g \leq P_g^*(\theta) \leq P_g(\theta) < 1 & ; \text{if} \quad \theta > \theta_0 \end{cases}$$

where

(7.11) $$\theta_0 = P_g^{-1}(c_g) \quad ,$$

provided that c_g is greater than zero. This result is quite natural, since it is reasonably assumed that the probability of success in solving item g will decrease by random guessing if the one attained by the due cognitive process is higher than the one attained by random guessing, and it will increase by random guessing if the latter probability is higher than the former. If we assume that the asymptotes of $d_g(\theta)$ in negative and positive directions be unity and zero, respectively, we will obtain c_g and unity as the lower and upper asymptotes of $P_g^*(\theta)$. Figure 7-1 presents two examples of the operating characteristic given by (7.8) where c_g is 0.2 , using two different $d_g(\theta)$'s . Note that there is a *dip* on the lower part of the curves for $P_g^*(\theta)$. These two $d_g(\theta)$'s are identical for the lower levels of θ , but differ on the upper levels, with the upper asymptotes 0.0 and 0.1 , respectively. In these examples, therefore, the upper asymptote of $P_g^*(\theta)$ is unity in the first example, and 0.92 in the second, i.e., the conditional probability for the correct answer never approaches unity however high the ability may be.

If $d_g(\theta)$ is differentiable, $P_g^*(\theta)$ is also differentiable, and from (7.8) we have

(7.12) $$\frac{\partial}{\partial\theta} P_g^*(\theta) = [1 - d_g(\theta)] \frac{\partial}{\partial\theta} P_g(\theta) + [c_g - P_g(\theta)] \frac{\partial}{\partial\theta} d_g(\theta) \quad .$$

Thus it is obvious that $P_g^*(\theta)$ is strictly increasing in θ for the range $\theta \geq \theta_0$, if, and only if, $d_g(\theta)$ is less than unity for the range of θ satisfying $\theta \geq \theta_0$. Thus in this case $P_g^*(\theta)$ is non-decreasing in θ throughout its whole range. In general, $P_g^*(\theta)$ equals c_g and presents a horizontal line as far as $d_g(\theta)$ is unity, and then increases for the rest of the range as θ increases.

As for the range expressed by $\theta \leq \theta_0$, $P_g^*(\theta)$ equals c_g regardless of the value of $P_g(\theta)$ for the values of θ for which $d_g(\theta)$ is unity, and is some positive value less than c_g otherwise. If $d_g(\theta)$ is unity throughout this range of θ , $P_g^*(\theta)$ presents a horisontal line for this range. If $d_g(\theta)$ is unity for the negative extreme value of θ , but $d_g(\theta)$ takes on some values less than unity for a subset of

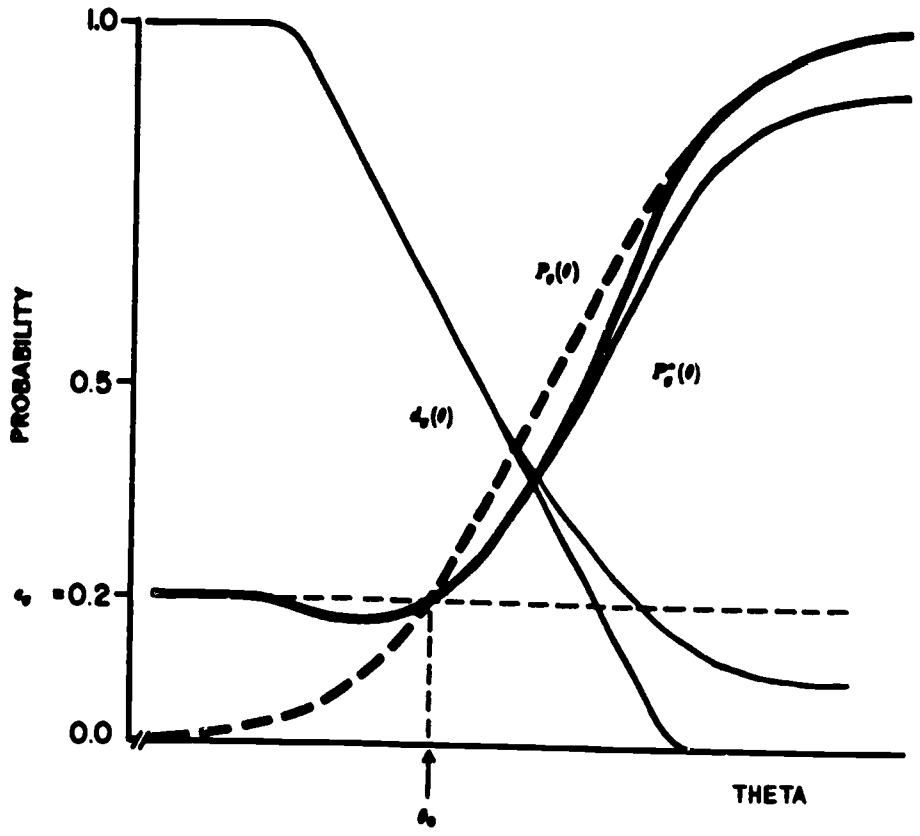

FIGURE 7-1

Relationships among $P_g(\theta)$, $d_g(\theta)$ and $P_g^*(\theta)$ Using Two Different $d_g(\theta)$'s.

θ of this range, $P_g^*(\theta)$ has at least one local minimum. If $d_g(\theta)$ is less than unity for the negative extreme value of θ, $P_g^*(\theta)$ can be strictly increasing in θ, non-decreasing, or have *one or more local minima*, in accordance with the functional formula for $d_g(\theta)$.

It is obvious that any operating characteristic having local minima does not satisfy the unique maximum condition (Samejima, 1969, 1972), and neither does the one whose first derivative equals zero at some value of θ. In the case of $P_g^*(\theta)$ defined by (7.8) we can prove that, in general, it *does not* satisfy the unique maximum condition, even if it is strictly increasing in θ. (For proof, see Samejima, 1968.)

Two characteristics of the model represented by (7.8) are that it allows *dips*, and also a smaller value than unity for the upper asymptote of the operating characteristic of the correct answer, as Figure 7-1 illustrates. In these examples, there is only one *dip* on the lower level of θ. There can be more than one, however, and an example is presented elsewhere (Samejima, 1968). In many cases the model may describe the real operating characteristic of the correct answer more closely than the three-parameter model.

It has been reported by several researchers that they have come across estimated operating characteristics of correct answers *that do not converge to unity, but to some other values less than unity*. Note that the general model described above can handle such situations, although most of the other models proposed by different researchers so far cannot.

We notice that neither (7.6) nor (7.8) explicitly takes into consideration the influences of separate

distractors. Suppose an examinee A has chosen to solve item g by reasoning, i.e., without guessing, and has reached an answer which is not correct. Suppose, further, that this specified response is *not* given as an alternative answer to this item. Then either he will decide to give an answer by guessing, or he will try to solve the item by reasoning all over again. To account for these possibilities, we would have to give practically all the different plausible responses to item g as its alternatives, which is practically impossible, since the number of alternative answers is more or less restricted. In contrast to this, it is interesting to note that the psychological hypothesis behind the three-parameter logistic model may be more realistic in the case where no very plausible responses except for the correct answer to item g are given as its alternative answers. Thus, even if an examinee has reached a specified plausible response other than the correct answer, he may turn to random guessing simply because he cannot find that specified answer among the alternatives. Such a situation has another serious problem, however, since it is likely for an examinee who is highly alternative-oriented to choose the correct answer without much reasoning or guessing, simply because the other alternatives are too ridiculous to be the answer to the item. As the result, the operating characteristic of the correct answer may be *deformed* so that it has a lower difficulty and less discriminating power. Plausible answers as *distractors* are necessary as alternatives in order not to destroy the nature of the item.

It is conceivable that the plausibilities of the alternatives attached to item g other than the correct answer will be one of the factors affecting the probability of random guessing in the multiple-choice situation. For this reason, here we shall suppose that an examinee will try to solve the item following proper cognitive processes at the beginning, and only in the case where he has reached an answer which is not given as an alternative, or where he has failed to find any answer at all, he will guess.

Let k_g or h_g denote a specified response to item g which is given as an alternative, including the correct answer, and $P_{k_g}(\theta)$ or $P_{h_g}(\theta)$ be its operating characteristic in the free-response situation. It may reasonably be assumed that $\sum_{k_g} P_{k_g}(\theta)$ is less than or equal to unity for any fixed value of θ. Let $P^*_{k_g}(\theta)$ or $P^*_{h_g}(\theta)$ denote the operating characteristic of a specified alternative k_g or h_g in the multiple-choice situation, and c_{k_g} or c_{h_g} be the probability of choosing k_g or h_g by guessing, which satisfies

$$\sum_{k_g} c_{k_g} = 1 \quad . \tag{7.13}$$

Thus we can write

$$P^*_{k_g}(\theta) = P_{k_g}(\theta) + [1 - \sum_{h_g} P_{h_g}(\theta)]\, c_{k_g} \tag{7.14}$$

for any k_g, and, by using the notation for the correct answer as we did in the previous sections, we obtain

$$P^*_g(\theta) = P_g(\theta) + [1 - \sum_{h_g} P_{h_g}(\theta)]\, c_g \quad . \tag{7.15}$$

It is worth noting that we have specified not only the operating characteristic of the correct answer in the multiple-choice situation, but also of each distractor. The utility of the operating characteristic of each wrong alternative answer in the estimation of an examinee's ability, as well as the one of the correct response, is suggested, and this is a feature of the present discussion.

It has been made clear that, in general, $P^*_g(\theta)$ *does not* satisfy the unique maximum condition regardless of the functional formulae for the plausibility curves of the distractors. As for the alternatives other than the correct answer, it can easily be shown that, in general, $P^*_{k_g}(\theta)$ does not satisfy the unique maximum condition (cf. Samejima, 1968, 1979b).

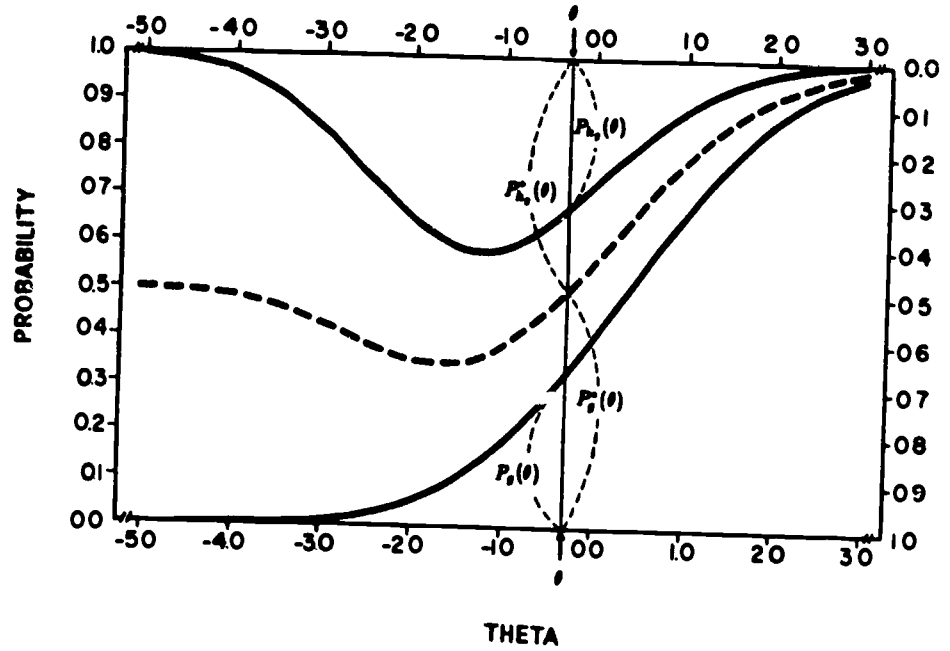

FIGURE 7-2

Operating Characteristic of the Correct Answer in the Free-Response Situation (Solid Line) and in the Multiple-Choice Situation (Dashed Line), in the Case Where Only Two Alternatives Are Given; Also the Operating Characteristic of the Other Alternative in the Free-Response Situation (Solid Line) Is Plotted from the Ceiling; $c_g = c_{k_g} = 0.5$.

For the purpose of illustration, Figure 7-2 presents a simple example in which only two alternatives, the correct answer and one incorrect response, are given. In this example, $P^*_{k_g}(\theta)$ for the wrong answer is drawn from the ceiling in order to make the picture visibly understandable. A normal ogive function given by

$$P_g(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_g(\theta - b_g)} \exp\{-u^2/2\}\, du \tag{7.16}$$

with $a_g = 1/1.48$ and $b_g = 0.36$ is used as the operating characteristic of the correct answer, and the same formula is applied for $U_{k_g}(\theta)$ and $M_{k_g}(\theta)$ for the incorrect response. The corresponding values of the parameters are $(1/1.23)$ and -1.84 for $M_{k_g}(\theta)$, and $(1/1.51)$ and -0.83 for $U_{k_g}(\theta)$. The value of c_g , as well as that of c_{k_g} for the incorrect answer, is 0.5 .

It is obvious from the above observations and discussion that these are the fundamental philosophies which led to the proposal of the new family of models for the multiple-choice test item (Samejima, 1979b). These philosophies will provide us with the idea of content-based observation of informative distractors and strategies of writing test items, which will be proposed in a later section. The general model described here is called Informative Distractor Model, in contrast with the Equivalent Distractor Model, to which the three-parameter model represented by (7.6) belongs (cf. Samejima, 1979b).

[VII.2] Effect of Noise in the Three-Parameter Logistic Model and the Meanings of the Difficulty and Discrimination Parameters

It is still a common procedure among researchers to adopt the three-parameter logistic model, which is represented by (3.11) in Section 3.2, for their multiple-choice test items and compare the resulting estimated discrimination parameters, or the difficulty parameters, across different items. An important fact that is overlooked is that *this is not legitimate*, for the addition of the third parameter c_g makes the other two item parameters lose their original meanings. If $a_g = 1.00$ and $c_g = 0.25$ in the three-parameter logistic model, for example, this corresponds to $a_g = 0.75$ in the logistic model in the maximum discrimination power. If, in addition to these parameter values, $b_g = 0.00$, then the difficulty level for the three-parameter logistic model defined as the level of θ at which chances for success are 0.5 is -0.4077336, i.e., substantially lower than 0.00.

In general, we can write

$$(7.17) \qquad \begin{cases} \alpha_g = (1 - c_g)\, a_g \\ \beta_g = b_g + (Da_g)^{-1} \log (1 - 2c_g) \ , \end{cases}$$

where α_g denotes the actual discrimination power and β_g is the actual difficulty level in the three-parameter logistic model. As we can see in (7.17), the effect of the third parameter c_g can be substantial, both on the discrimination power α_g and on the difficulty index β_g. Thus the simple comparison of the values of a_g for two or more test items having different values of the lower asymptote c_g is illegitimate and can be harmful, for the factor $(1 - c_g)$ may affect the value of α_g, the real discrimination power, substantially. As for the difficulty index, since the second term on the right hand side of the second equation of (7.17) is always negative for $0 < c_g < 0.5$, this term represents the amount of decrement of the difficulty level. Note that as c_g tends to 0.5, β_g approaches negative infinity! (If $c_g \geq 0.5$ then β_g does not even exist.) The illegitimacy of, and the danger in, comparing b_g's across two or more test items having different lower asymptotes c_g is even more obvious for the difficulty index.

It is obvious from theory that in both the logistic and the three-parameter logistic models the derivative of the operating characteristic of the correct answer is highest at $\theta = b_g$. Actually, the derivatives are: $Da_g/4$ and $(1 - c_g)Da_g/4$, respectively. The ratio of this maximal slope between the three-parameter logistic model and the logistic model is $(1 - c_g)$, which equals 0.75 when $c_g = 0.25$, and is as low as 0.50 when $c_g = 0.50$. The corresponding ratio between the three-parameter logistic model and the normal ogive model is approximately $0.938687718(1 - c_g)$, which is a little less than $(1 - c_g)$.

Figure 7-3 illustrates that several sets of substantially different parameter values in the three-parameter logistic model can produce very similar operating characteristics of the correct answer. We can tell that the differences in the values of the discrimination and difficulty parameters for these items are substantial, and yet the resulting curves are very close to each other for a wide range of θ. *Simple comparison of the two estimated discrimination parameters is illegitimate*, therefore, when the estimated guessing parameters prove to be different from each other, as is usually the case with actual data. Since the estimation of the third parameter c_g tends to be most inaccurate, this example indicates the danger in direct comparisons of the estimated discrimination parameters, and also the estimated difficulty parameters, across the items.

In most cases the estimated guessing parameter of a multiple-choice test item provides us with some other value than the reciprocal of the number of the alternative answers. It is reported that in some cases the estimated c_g takes on quite high values (cf. Lord, 1980, Section 2.2). These phenomena suggest that the philosophy behind the model is unrealistic. Researchers using the three-parameter logistic model argue, however, that it still is a convenient approximation to real operating characteristics of

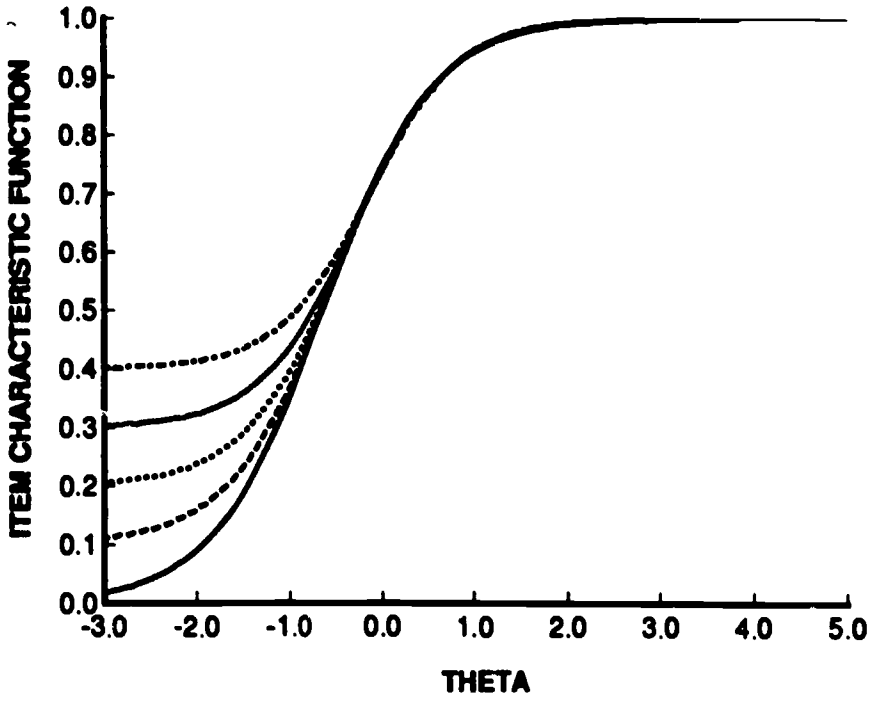

FIGURE 7-3

Examples of the Operating Characteristics of the Correct Answer in the Three-Parameter Logistic Model (Dotted Lines), Together with the One in the Logistic Model with $a_g = 1.00$ and $b_g = -0.64$ (Solid Line). The Parameters for the Four Functions in the Order of a_g, b_g and c_g are: 1.05, -0.52, 0.10; 1.10, -0.40, 0.20; 1.15, -0.27, 0.30; 1.20, -0.13, 0.40; Respectively.

correct answers, because of its simplicity in mathematics. In a way it is true. The effective use of the three-parameter model cannot be realized, however, unless we know the problems attributed to the model, and use the model in such a way that these weaknesses will not cause too much noise and inefficiency.

Investigation of the problems encountered when we apply the three-parameter logistic model to the data which actually follow the normal ogive model was made earlier (Samejima, 1984b). The data used in the study are simulated data for two samples of 500 and 2,000 hypothetical examinees, respectively, sampled from the uniform ability distribution for the interval of θ, (-2.5, 2.5). In order to investigate the effect of the number of test items on the resultant estimated parameters obtained by Logist 5, we used: 1) Ten Item Test and 2) Thirty-Five Item Test, both of which consist of binary items following the normal ogive model. The response pattern for each hypothetical subject was produced by the Monte Carlo Method. Combining these two hypothetical tests, we observed the results of: 3) Forty-Five Item Test, and, in addition, we observed the results of rather artificially created: 4) Eighty Item Test (cf. Samejima, 1984b).

These results suggest that there exists a substantial effect of the assumed third parameter, c_g, on the other two estimated item parameters, if the estimation is made by *molding* the operating characteristic of the correct answer into that of the three-parameter logistic model, when actually it follows the normal ogive model. This effect appears to be stronger on the estimated discrimination parameter than on the estimated difficulty parameter. In order to amend these enhancements, the discrimination shrinkage

factor and the difficulty reduction index were proposed (Samejima, 1984b) by formulae (7.19) and (7.21), respectively.

$$a_g^* = \varsigma(c_g^*)\, a_g \quad . \tag{7.18}$$

$$\varsigma(c_g^*) = -\log(1-2c_g^*)\,\log(1+c_g^*) - \log(1-c_g^*)^{-1} \quad . \tag{7.19}$$

$$b_g^* = b_g + \xi(c_g^* \mid a_g) \quad . \tag{7.20}$$

$$\xi(c_g^* \mid a_g) = (Da_g)^{-1}\log(1+c_g^*) - \log(1-c_g^*) \quad . \tag{7.21}$$

In these formulae, a_g^*, b_g^*, and c_g^* indicate the estimated item discrimination, difficulty and guessing parameters when the three-parameter logistic model is assumed, respectively. Some resulting estimated operating characteristics of the correct answer turned out to be disastrously different from the theorectical functions, especially when only ten binary test items were included. We find no substantial differences between the results of 500 Subject Case and 2,000 Subject Case, indicating that increasing the number of subjects from 500 to 2,000 does not provide us with a substantial gain.

It has been pointed out that the three-parameter logistic model does not satisfy the unique maximum condition for the likelihood function, and this topic has been thoroughly discussed (Samejima, 1973). The expected loss of item information for a fixed value of θ is given by

$$I_g(\theta) - I_g^*(\theta) = c_g D^2 a_g^2 [\{\psi_g(\theta)\}^2\{1-\psi_g(\theta)\}][c_g + (1-c_g)\psi_g(\theta)]^{-1} \quad , \tag{7.22}$$

where

$$\psi_g(\theta) = [1+\exp\{-Da_g((\theta)-b_g)\}]^{-1} \quad , \tag{7.23}$$

and $I_g(\theta)$ and $I_g^*(\theta)$ are the item information functions in the logistic and the three-parameter logistic models, respectively. We have for the critical value θ_g, below which the information provided by the correct answer to the item following the three-parameter logistic model assumes negative values

$$\theta_g = b_g + (2Da_g)^{-1}\log c_g \quad , \tag{7.24}$$

which is strictly increasing with the increase in the parameter value c_g, and also in a_g and in b_g. If, for example, $a_g = 1.00$ and $b_g = 0.00$, $\theta_g = -0.473364$ for $c_g = 0.20$, and $\theta_g = -0.407734$ for $c_g = 0.25$. They are considerably high values relative to b_g.

An important implication is that θ_g is the point of θ below which the existence of a unique maximum likelihood estimate is not assured for all the response patterns which include the correct answer to item g. Although this warning has been ignored by most researchers for many years, a recent research (Yen, Burket and Sykes, in press) points out this is happening much more often than people might think.

It has been pointed out (Samejima, 1979a, 1982a) that there is a certain constancy in the total amount of item information, regardless of the parameter values and of specific functional formulae for the operating characteristic of the correct answer. If, for example, the model belongs to Type A, i.e., the operating characteristic of the correct answer is monotone increasing with zero and unity as its lower and upper asymptotes, respectively, then the total area under the curve of the square root of the

item information function will equal π. If the model belongs to Type B, i.e., the same as Type A except that the lower asymptote of the operating characteristic of the correct answer is greater than zero, as is the case with the three-parameter logistic model, then the total area will become

$$\pi - 2\tan^{-1}[c_g(1-c_g)^{-1}]^{1/2} \quad , \tag{7.25}$$

with the second and last term as *the loss in the amount of total item information*. This last term is strictly a function of c_g. When $c_g = 0.20$, for example, the total amount of item information reduces, approximately, to 0.705π, and when $c_g = 0.25$ it is approximately equal to 0.667π. More observations concerning the effect of noise in the three-parameter logistic model have been made elsewhere (Samejima, 1982b).

As all the above observations indicate, the addition of the third parameter, c_g, to the logistic model creates many negative results. We have seen that these negative effects are greater for larger values of c_g. In using the three-parameter logistic model as an approximation to real operating characteristics, therefore, we need to take these facts into consideration. Among others, if we are in a situation where we can modify or revise our items, we must try to reduce the effect of noise coming from c_g as much as possible. Strategies of writing the multiple-choice test items must be considered accordingly.

[VII.3] Informative Distractors of the Multiple-Choice Test Item

So far most observations and discussion have been focused on theory. Applications of certain nonparametric methods of estimating the operating characteristics for some empirical data have revealed, however, that many multiple-choice test items do not follow the three-parameter model, nor do they follow the Equivalent Distractor Model in general, to which the three-parameter logistic model belongs. Those items can best be interpreted by the Informative Distractor Model.

Figure 7-4 presents an example of the set of operating characteristics of the four alternative answers to an item taken from the Level 11 Vocabulary Subtest of the Iowa Tests of Basic Skills (Samejima, 1984a), which was estimated by the Simple Sum Procedure of the Conditional P.D.F. Approach combined with the Normal Approach Method (cf. Section 6.1). We can see in this figure that each distractor has its own unique operating characteristic, or *plausibility function*, and also that the estimated operating characteristic of the correct answer is fairly close to the one in the normal ogive model, which is drawn by a solid line in the figure. This set of operating characteristics can better be represented by one of the family of models proposed for the multiple-choice test item, which was originated by the philosophy described in the preceding section and takes account of the unique information provided by each distractor as well as the effect of the examinees' random guessing behavior (cf. Samejima, 1979b). Figure 7-5 illustrates the operating characteristic of the correct answer in Model A. We can see that it is very close to the one in the normal ogive model which is drawn by a dotted line, except for the lower part of the curve, the conditional probability of success which is almost entirely caused by random guessing. In cases like this, *it will be wise to approximate the curve by the normal ogive function by discarding the item response in estimating lower ability*, since it provides us with nothing but noise, as was discussed in the preceding section.

Detailed observations for the plausibility functions of *distractors* are made elsewhere (Samejima, 1984a) for the forty-three items of the Level 11 Vocabulary Subtest of the Iowa Tests of Basic Skills. Similar *discoveries* have also been reported with respect to many ASVAB test items. In those results, it is clear that separate wrong answers given as alternatives provide us with differential information, which can be useful in ability estimation in the sense that it will substantially increase the accuracy of estimation.

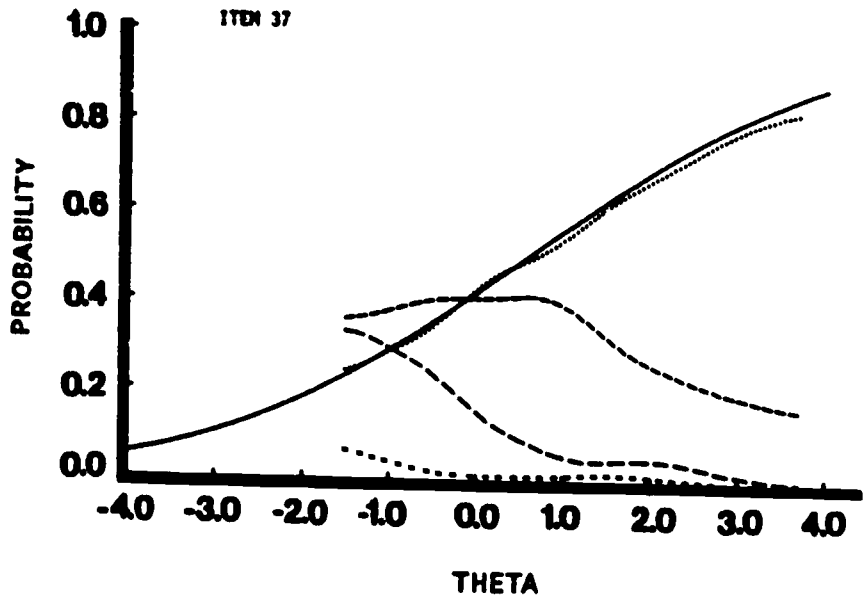

FIGURE 7-4

Example of the Estimated Operating Characteristics of the Correct Answer (Dotted Line) and of the Three Distractors (Dashed Lines) Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach Combined with the Normal Approach Method Together with the One for the Correct Answer Obtained by Assuming the Normal Ogive Model (Solid Line) Taken from the Level 11 Vocabulary Subtest of the Iowa Tests of Basic Skills.

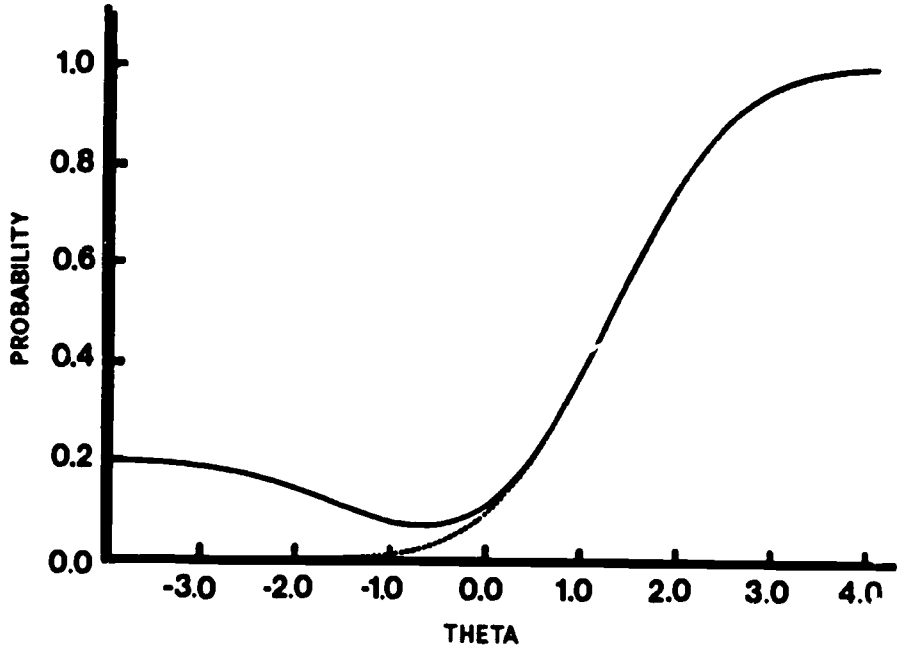

FIGURE 7-5

Example of the Operating Characteristic of the Correct Answer in Model A (Solid Line) Together with One in the Normal Ogive Model (Dotted Line).

[VII.4] Merits of the Nonparametric Approach for the Identification of Informative Distractors and for the Estimation of the Operating Characteristics of an Item

Methods and approaches developed for estimating the operating characteristics of discrete item responses without assuming any mathematical form (cf. Section 2.3; Samejima, 1981, 1990) enable us to find out whether or not a given incorrect alternative answer to a multiple-choice test item is informative in the sense that it contributes to the increment in the accuracy in the estimation of the individual's ability. Recently, the author proposed a new approach, which is called Differential Weight Procedure of the Conditional P.D.F. Approach, and which has been described in the preceding chapter. Although we need more research for improving the fitnesses further, those results obtained so far give us promises for success in identifying informative distractors and in estimating their operating characteristics.

Item analysis has a long history, starting from the classical proportion correct and item-test regression. In the context of latent trait models, the operating characteristics and the information functions have provided us with powerful tools. Now we can add the plausibility functions of the distractors to this category. By accurately identifying the configuration of the operating characteristics of the correct answer and the distractors, we shall be able to understand the characteristics of the item, its strengths and weaknesses. In this way modifications of the item can be done if necessary. Successful nonparametric methods of estimating the operating characteristics are essential, therefore, for this new, more informative approach to the item analysis.

[VII.5] Efficiency in Ability Estimation and Strategies of Writing Test Items

Observations and discussion made in the preceding sections give us much useful information as well as warnings. First of all, theoretical observations indicate that non-monotonicity of the operating characteristic of the correct answer to the multiple-choice test item is a natural consequence of theory. Secondly, it has been shown from several different angles that the third parameter, c_g, in the three-parameter model provides us with nothing but noise; the greater the value of c_g the more noise and inaccuracies in estimation it produces. Thirdly, it has been pointed out that, although it is still a common procedure for researchers to *mold* the operating characteristics of the correct answers of their multiple-choice test items into the three-parameter logistic model, some nonparametric methods applied to *empirical* data have revealed the non-monotonicity of the operating characteristic of the correct answer with many *actual* test items, as well as differential information provided by separate distractors. Fourthly, it has been pointed out that the nonparametric approach to the estimation of the operating characteristics of discrete responses has been successful enough to detect the non-monotonicity of the function when it exists, and to approximate their rather irregular curves fairly accurately.

With all these facts, *it is time to reconsider conventional strategies for item writing* and to propose new strategies.

The first thing we need to reconsider is the lack of sufficient interactions between theorists and people who write test items. It has been fairly common that: 1) a committee is organized for writing test items in a specified content area or domain and eventually produces a set of test items; 2) another group of people tests these items on a small sample of subjects, screens the items and then administers the selected items to larger groups of subjects. Item calibration is done on the second stage, assuming some model such as the three-parameter logistic model, etc. In most cases, there is practically no feedback from theorists to item writers. If we set a strategy that more interactions are made between the two groups of people so that the test items are revised and pilot tested with each interaction, we shall be able to improve the test, and the improvement will lead to efficiency in ability estimation.

The second thing we need to reconsider is the simpleminded avoidance of non-monotonicity of the

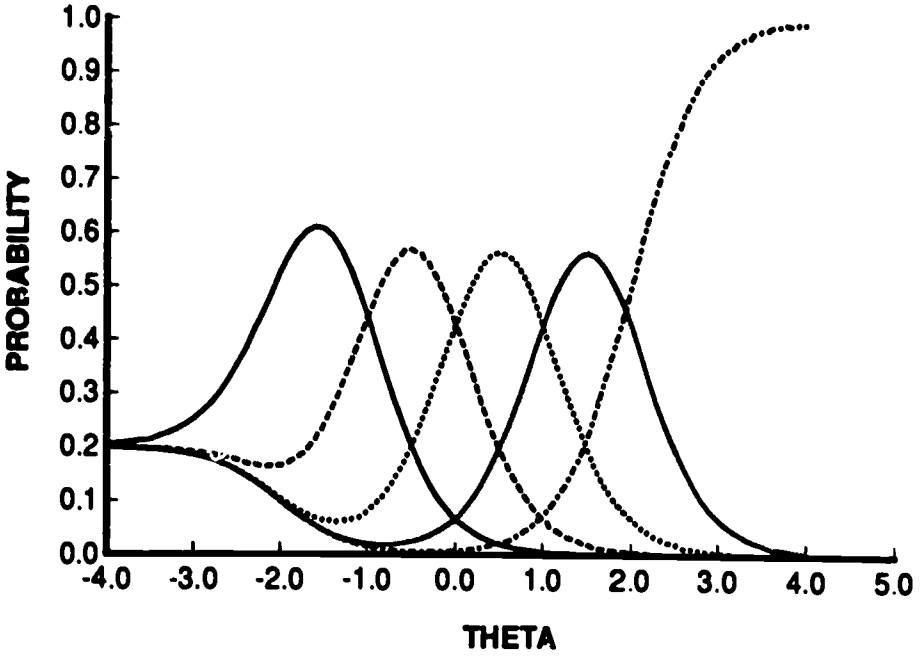

FIGURE 7-6

Operating Characteristics of the Five Alternative Answers of a Hypothetical Test Item Following Model B, with the Parameter Values: $a_g = 1.5$, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 0.0$, $b_4 = 1.0$ and $b_5 = 2.0$.

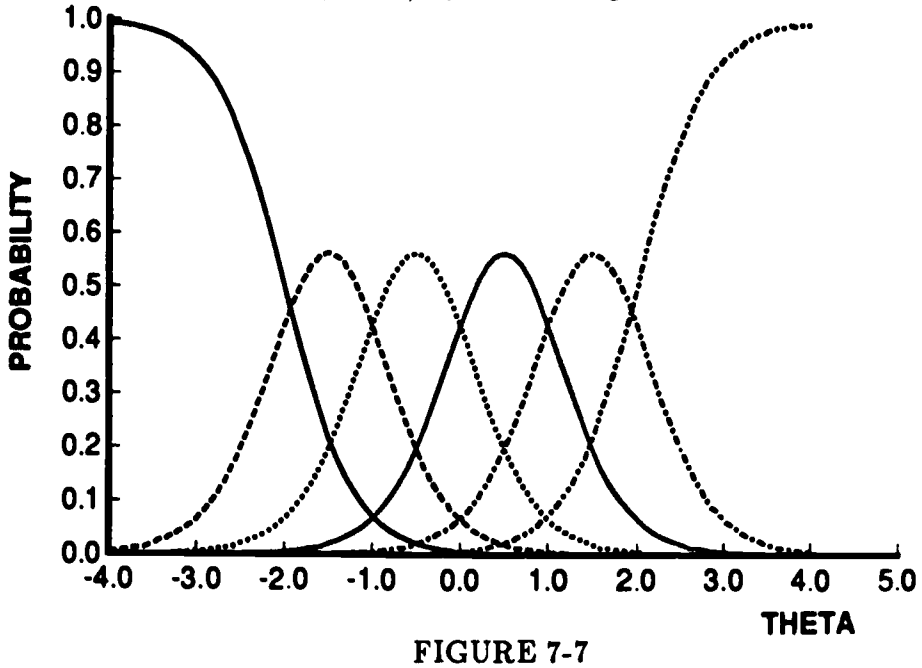

FIGURE 7-7

Operating Characteristics of the Five Alternative Answers of a Hypothetical Test Item in the Free-Response Situation Following the Logistic Model on the Graded Response Level, with the Parameter Values: $a_g = 1.5$, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 0.0$, $b_4 = 1.0$ and $b_5 = 2.0$.

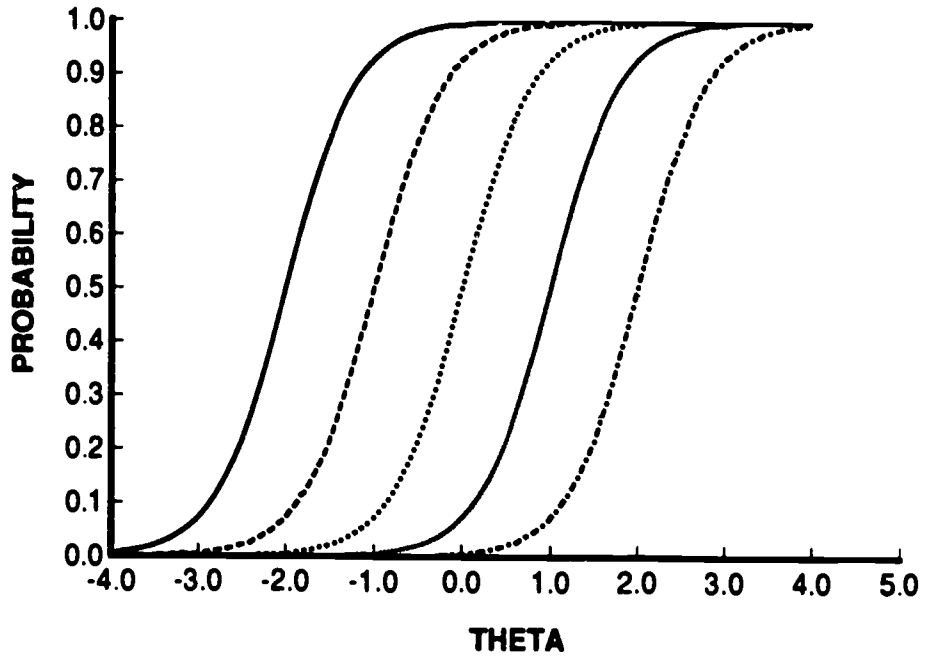

FIGURE 7-8

Operating Characteristics of the Correct Answer Obtained by the Five Different Redichotomisations of the Graded Test Item Following the Logistic Model, with the Discrimination Parameter, $a_g = 1.5$, and the Difficulty Parameters, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 0.0$, $b_4 = 1.0$ and $b_5 = 2.0$, Respectively.

operating characteristic of the correct answer. While it is not desirable for an item to have higher conditional probabilities of the correct answer on lower levels of ability than on h ‘ier levels, selecting alternative answers so that the *dips* of the operating characteristic of the correct answer be *smoothed out* will lead to a substantially large value of the lower asymptote of the operating characteristic in most ca.es. We must recall that even a small number like 0.2 as c_g in the three-parameter logistic model is a big nuisance, as was discussed in Section 7.2. Our strategy must be that we make the best use of those *dips*, instead of avoiding them.

Figure 7-6 presents the operating characteristics of the five alternative answers of a hypothesized test item following Model B (Samejima, 1979b), with the parameter values: $a_g = 1.50$, $b_1 = -2.00$, $b_2 = -1.00$, $b_3 = 0.00$, $b_4 = 1.00$ and $b_5 = 2.00$. The subscript for each of the five difficulty parameters indicates the order of *easiness* for the examinee to be attracted to the plausibility of each alternative answer, so that, in this example, b_5 indicates the difficulty parameter of the correct answer. We can see in this figure that a practical monotonicity exists for the operating characteristic of the correct answer for the range of θ , $(-0.5, \infty)$, and, more importantly, within this range of θ its lower asymptote is very close to zero, i.e., the nuisance caused by the non-zero lower asymptote will be gone as far as we administer the item to populations of subjects whose ability distributes on higher levels than $\theta = -0.5$.

These operating characteristics of the five alternative answers in Figure 7-6 are originated from those in the logistic model on the graded response level (Samejima, 1969, 1972) with the same parameter values (cf. Samejima, 1979b). Figure 7-7 presents the corresponding set of operating characteristics of the correct answers in the logistic model. We notice there is an additional strictly decreasing curve in this figure. This curve represents the conditional probability, given θ , that the examinee does not find

attractiveness in any alternative answers. In Model B, these people are assumed to guess randomly, so in Figure 7-6 this curve does not exist, and the conditional probability is evenly distributed among the five alternative answers to account for the rises in their operating characteristics at lower levels of θ .

Figure 7-8 presents the operating characteristics of the correct answer following the logistic model on the dichotomous response level, which are obtained by the five different redichotomisations of the graded test item exemplified in Figure 7-7. In these functions, $a_g = 1.5$ is the common discrimination parameter, and the difficulty parameters are: $b_g = -2.0, -1.0, 0.0, 1.0, 2.0$, respectively. This is the starting point of the graded response model, which leads to the operating characteristics illustrated in Figure 7-7 (cf. Samejima, 1969, 1972).

Suppose that two alternative answers which attract examinees of low levels of θ are replaced, and the revised item has $b_1 = -3.0$ and $b_2 = -1.5$, respectively. In this situation, the operating characteristics of the correct answer obtained by the first two redichotomisations are changed. Figure 7-9 presents the set of operating characteristics for this revised test item following Model B. In this figure we can see that the operating characteristic of the correct answer is practically strictly increasing within the range of θ , $(-1.7, \infty)$, and the *pseudo* lower asymptote of the operating characteristic within this range of θ is still very close to zero.

A big gain resulting from this revision is the fact that the lower endpoint of the interval of θ in which the operating characteristic of the correct answer is practically monotonic has substantially shifted to the negative direction, while still keeping its lower asymptote practically zero. Thus we can avoid the noise coming from the lower asymptote even if we administer the item to populations of examinees whose ability distributions are located on lower levels of θ . In other words, without sacrificing the accuracy of ability estimation, the utility of the item has been substantially enhanced by this revision.

The above example suggests the following strategy.

(1) If the nonparametrically estimated operating characteristic of the correct answer to an item provides us with a relatively high value of θ below which monotonicity does not exist, then change the set of distractors to include one or more wrong answers that attract examinees of very low levels of ability.

It may sound difficult to do in practice. If we pay attention to actually used multiple-choice test items, however, we will come across many wrong alternative answers that are attracting examinees of very low levels of ability. To give an example, the author has come across an arithmetic item asking for the area of a rectangle. A substantial number of seventh graders chose the wrong alternative answer which equals the sum of the two sides of the rectangle of different lengths! It is obvious that those who did not understand how to obtain the area of a rectangle at all chose this alternative answer.

Another consideration which is important in writing test items is to keep the *pseudo* lower asymptote of the operating characteristic of the correct answer *close enough to zero*, as is the case with the above example. This has a great deal to do with the discrimination powers of the alternative answers, as well as the configuration of the plausibility functions. Figure 7-10 presents the set of operating characteristics corresponding to Figure 7-6, by changing the discrimination parameter from $a_g = 1.5$ to $a_g = 1.0$, while keeping the five difficulty parameters unchanged. If we compare Figure 7-10 with Figure 7-6, we can see a substantial enhancement of the *pseudo* lower asymptote within the interval of θ , $(-0.5, \infty)$, i.e., the nuisance has been increased by the change in the discrimination parameter.

This suggests the second strategy:

(2) If possible, try to include distractors whose estimated operating characteristics are *steep*, while keeping the differential configuration of these functions as suggested in (1).

So far our strategies have been focused upon producing an informative operating characteristic of

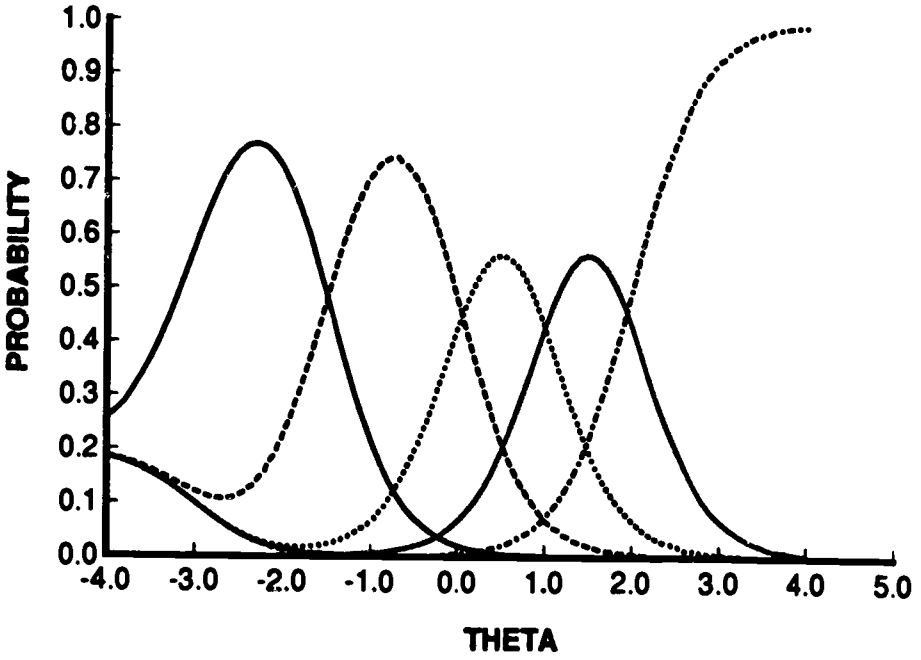

FIGURE 7-9

Operating Characteristics of the Five Alternative Answers of a Hypothetical Test Item Following Model B, with the Parameter Values: $a_g = 1.5$, $b_1 = -3.0$, $b_2 = -1.5$, $b_3 = 0.0$, $b_4 = 1.0$ and $b_5 = 2.0$.

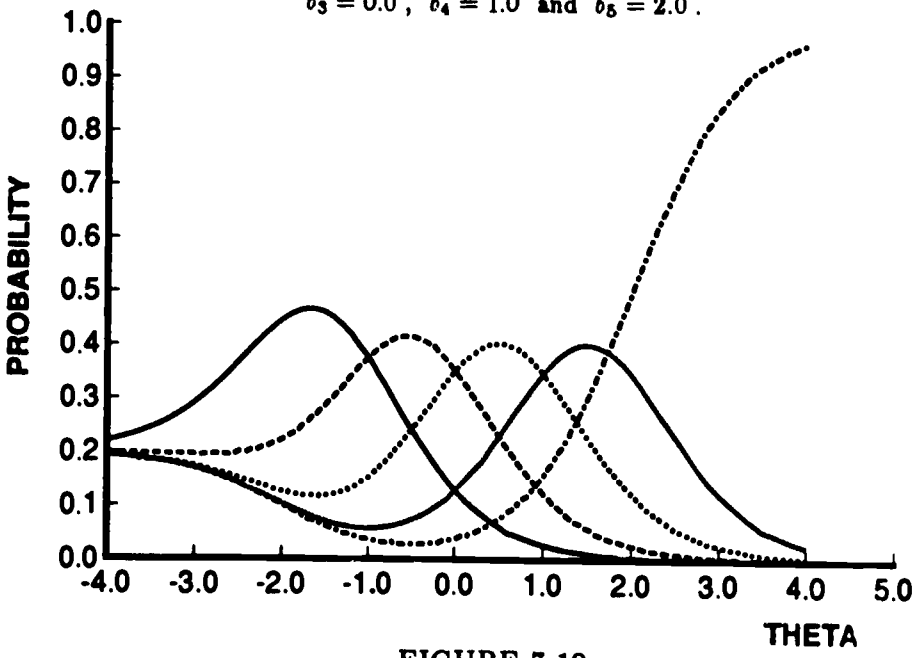

FIGURE 7-10

Operating Characteristics of the Five Alternative Answers of a Hypothetical Test Item Following Model B, with the Parameter Values: $a_g = 1.0$, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 0.0$, $b_4 = 1.0$ and $b_5 = 2.0$.

the correct answer. We notice, however, that *these strategies will also provide us with distractors which provide us with differential information.* This implies that approximation of the nonparametrically estimated operating characteristics of one or more alternative answers by some mathematical formulae will enable us to use this additional differential information in ability estimation. This *posterior parameterization* of the non-parametrically estimated operating characteristics of distractors will certainly lead us to increased accuracy and efficiency in ability measurement.

[VII.6] Discussion and Conclusions

In this chapter, the shortages of the conventional way of handling the multiple-choice test have been summarised, and also theories and methodologies that can be applied for a better handling of the multiple-choice test item have been described; some empirical facts have been introduced to support the theoretical observations; finally, new strategies of item writing have been proposed which will reduce *noise* and lead to more efficient ability estimation.

In spite of many controversies against the multiple-choice test, because of its economy in scoring it has been, and still is, very popular among people of psychological and educational measurement. Fortunately, theorists in mathematical psychology have developed many new ideas and methodologies in the past couple of decades that can improve the way of handling the multiple-choice test. Nonparametric approach in estimating the operating characteristic is one of them. Also the rapid progress in electronic technologies has made it possible to materialise these results of theories and methodologies in practical situations. Today, we are in a position to take advantage of all these accomplishments.

References

[1] Bock, R. D. and Aitkin, M. Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 1981, 443-459.

[2] Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

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

[4] Samejima, F. Application of the graded response model to the nominal response and multiple-choice situations. *UNC Psychometric Laboratory Report, 63,* 1968.

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

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

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

[8] Samejima, F. Constant information model: A new, promising item characteristic function. *ONR/RR-79-1,* 1979a.

[9] Samejima, F. A new family of models for the multiple-choice item. *ONR/RR-79-4,* 1979b.

[10] Samejima, F. Final Report: Efficient methods of estimating the operating characteristic of item response categories and challenge to a new model for the multiple-choice item. *Final Report of N00014-77-C-0360,* Office of Naval Research, 1981.

[11] Samejima, F. Information loss caused by noise in models for dichotomous items. *ONR/RR-82-1,* 1982a.

[12] Samejima, F. Effect of noise in the three-parameter logistic model. *ONR/RR-82-2,* 1982b.

[13] Samejima, F. Plausibility functions of Iowa Vocabulary Test items estimated by the Simple Sum Procedure of the Conditional P.D.F. Approach. *ONR/RR-84-1*, 1984a.

[14] Samejima, F. Results of item parameter estimation using Logist 5 on simulated data. *ONR/RR-84-9*, 1984b.

[15] Samejima, F. Differential Weight Procedure of the Conditional P.D.F. Approach for estimating the operating characteristics of discrete item responses. *ONR/RR-90-4*, 1990.

[16] Wingersky, M. S., Barton, M. A., and Lord, F. M. *Logist user's guide*. Princeton: Educational Testing Service, 1982.

[17] Yen, W. M., Burket, G. R., and Sykes, R. C. Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika*, in press.

VIII Efficient Computerized Adaptive Testing

In the previous chapters, various research findings obtained in the present research period have been introduced and discussed. All of these results are beneficial for computerized adaptive testing, especially in increasing its efficiency. This chapter will summarise observations as to how these findings and developments can be applied in computerised adaptive testing.

[VIII.1] Validity Measures Tailoring a Sequential Subset of Items for an Individual

The item information function, $I_g(\theta)$, has been used in the computerized adaptive testing in selecting an optimal item to *tailor* a sequential subtest of items for an individual examinee out of the prearranged itempool. A procedure may be to let the computer choose an item having the highest value of $I_g(\theta)$ at the current estimated value of θ for the individual examinee, which is based upon his responses to the items that have already been presented to him in sequence, out of the set of remaining items in the itempool.

We notice from (5.6) or (5.8) in Section 5.2 that this procedure is also supported from the standpoint of maximising the criterion-oriented validity, for the item which provides us with the greatest item information $I_g(\theta)$ among all the available items in the itempool also gives the greatest values of $I_g^*(\zeta)$ and its square root, at any fixed value of θ .

[VIII.2] Use of the Modifications of the Test Information Function in Stopping Rules

It is a big advantage of the modern mental test theory over classical mental test theory that the standard error of estimation can *locally* be defined by means of $[I(\theta)]^{-1/2}$, which *does not depend upon the population of examinees,* but is solely a property of the test itself. Using this characteristic, it has been observed (Samejima, 1977) that in computerized adaptive testing the amount of test information can be used effectively in the stopping rule indicating, locally, the desirable accuracy of estimation of the examinee's ability, provided that our itempool contains a large number of items whose difficulty levels distribute widely over the range of θ of interest. A procedure may be to terminate the presentation of a new item out of the itempool to the individual examinee when $I(\theta)$ has reached an a priori set amount at the current value of his estimated θ .

We notice that, in general, for the stopping rule in computerized adaptive testing the modified test information functions, $\Upsilon(\theta)$ and $\Xi(\theta)$, will serve better than the original $I(\theta)$, for in many

practical situations our itempool is more or less limited. In particular, it is usual that there are not so many optimal items for examinees whose ability levels are close to the upper or the lower end of the configuration of the difficulty parameters of the items in the itempool. In such a case, even if the amount of test information has reached a certain criterion level, it does not mean that their ability levels are estimated with the same accuracy as those of individuals of intermediate ability levels, as was pointed out in Chapter 3. Since, taking the MLE bias function into consideration, the two modified test information functions, $\Upsilon(\theta)$ and $\Xi(\theta)$, are based upon a more meaningful minimum bound of the conditional variance and upon a minimum bound of the mean squared error of the maximum likelihood estimator, respectively, they will be effectively used as the replacement of $I(\theta)$ in stopping rules of computerized adaptive testing.

The test information function $I(\theta)$ and its two modification formulae, $\Upsilon(\theta)$ and $\Xi(\theta)$, are likely to be the ones exemplified in the lower graph of Figure 3-5 for an individual examinee in the process of adaptive testing provided that the program for the test is written well. We should expect visible differences between the results obtained by using $I(\theta)$ and by using one of its modification formulae, therefore, especially for subjects whose ability levels are close to the upper or lower end of the ability interval of interest. It is expected that these individuals will be required to take more test items in order to make the accuracy of the estimation of θ comparable to that of examinees of intermediate ability levels: a fact that could not have been disclosed without $\Upsilon(\theta)$ and $\Xi(\theta)$.

We need to investigate this topic in the future, specifying the amount of improvement with simulated and empirical data collected in computerized adaptive testing.

[VIII.3] Use of Test Validity Measures in Stopping Rules

When we have a specific criterion variable γ in mind, it is justified to use an a priori set value of $I^*(\varsigma)$ instead of $I(\theta)$ in the stopping rule of computerized adaptive testing. In so doing, we can obtain the value of $I(\theta)$ corresponding to the a priori set value of $I^*(\varsigma)$ for each θ, through the formula

$$I(\theta) = I^*(\varsigma)\,(\frac{\partial \varsigma}{\partial \theta})^2 \quad , \tag{8.1}$$

which is obtained from (5.9) in Section 5.2. Thus it is easy to have the computer handle this situation, provided that we know the functional formula for $\varsigma(\theta)$.

We notice that the test validity measures proposed in the present research (cf. Chapter 5) can be modified, if we replace the test information function $I(\theta)$ by one of its modification formulae, $\Upsilon(\theta)$ and $\Xi(\theta)$, which have also been proposed in the present research (cf. Chapter 3). This will be pursued in the future, when the characteristics of these two modified test information functions have further been pursued and clarified. It is quite possible that the new test validity measures can effectively be used in stopping rules of computerized adaptive testing.

[VIII.4] Prediction of the Reliability Coefficient for a Specific Population of Examinees in Computerized Adaptive Testing

It has also been observed (Samejima, 1977) 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 two conditions 1) and 2) described in Section 4.2 are met. In such a case, we can write

$$Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - E[\{I_{(1)}(\theta)\}^{-1}]][Var.(\hat{\theta}_1)\{Var.(\hat{\theta}_1) - E[\{I_{(1)}(\theta)\}^{-1}] \tag{8.2}$$

$$+ E[\{I_{(2)}(\theta)\}^{-1}]\}]^{-1/2} ,$$

where $I_{(1)}(\theta)$ and $I_{(2)}(\theta)$ 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

$$(8.3) \qquad Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - \sigma_1^2][Var.(\hat{\theta}_1)\{Var.(\hat{\theta}_1) - \sigma_1^2 + \sigma_2^2\}]^{-1/2} ,$$

where σ_1^2 and σ_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

$$(8.4) \qquad Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - \sigma^2][Var.(\hat{\theta}_1)]^{-1} ,$$

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

Also in computerised adaptive testing, either $\Upsilon(\theta)$ or $\Xi(\theta)$ can be used as the stopping rule in place of the test information function $I(\theta)$, and we can revise (8.2) into the forms

$$(8.5) \qquad Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - E[\{\Upsilon_{(1)}(\theta)\}^{-1}]][Var.(\hat{\theta}_1)\{Var.(\hat{\theta}_1) - E[\{\Upsilon_{(1)}(\theta)\}^{-1}]$$

$$+ E[\{\Upsilon_{(2)}(\theta)\}^{-1}]\}]^{-1/2} ,$$

and

$$(8.6) \qquad Corr.(\hat{\theta}_1, \hat{\theta}_2) = [Var.(\hat{\theta}_1) - E[\{\Xi_{(1)}(\theta)\}^{-1}]][Var.(\hat{\theta}_1)\{Var.(\hat{\theta}_1) - E[\{\Xi_{(1)}(\theta)\}^{-1}]$$

$$+ E[\{\Xi_{(2)}(\theta)\}^{-1}]\}]^{-1/2} ,$$

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

[VIII.5] Differential Weight Procedure for Item Analysis and for On-Line Item Calibration

It is obvious that *item analysis* in the true sense of the word starts from the accurate estimation of the operating characteristics of the item responses. Thus the nonparametric estimation of the operating characteristic offers a great deal of information about an item, when it is successful. In this sense we can say that the Differential Weight Procedure of the Conditional P.D.F. Approach (cf. Chapter 6) provides us with promise for the successful item analysis in general.

For the success in adaptive testing, it is essential to create a good initial itempool. Differential Weight Procedure can effectively be used in selecting appropriate test items for the itempool, applied repeatedly in pilot studies.

Differential Weight Procedure will especially be useful for the on-line item calibration in computerised adaptive testing. When we use an adaptive test, it is necessary to discard certain test items from our itempool after they have been administered too frequently, or too seldom, and replace them by new test items. In so doing, we need to on-line calibrate these new test items, and successful nonparametric estimation methods adjusted to this situation will be most valuable in order to *discover* the operating characteristics of these new test items.

Many computer programs have been written in the precent research, in order to materialise this new method, and to put the theory and methodologies in practice. In developing this method further, it will be the focus of research to pursue methodologies for estimating differential weight functions under different circumstances. It should also be noted that we need to develop efficient computer programs for smoothing out the irregularities of the differential weight function whenever it is needed.

Once the operating characteristics of the test items have been *discovered,* however, it will be wise to search for appropriate mathematical forms in order to mathematically simplify them by parameterisation. In so doing, observations and mathematical models introduced in Chapter 7 will be useful, especially in dealing with non-monotonic operating characteristics or those which are strictly increasing but converging to some values less than unity.

[VIII.6] Use of Informative Distractors

One of the future directions of the computerised adaptive testing will be the use of information coming from the *distractors* of the multiple-choice test item, as well as from the correct answer. This will certainly increase the item information both locally and in total, and, as the result, the estimation of the individual examinee's ability will become more efficient.

For this reason, an accurate estimation of the *plausibility functions* of the distractors of multiple-choice test items becomes very important for the future of computerised adaptive testing. In this context, again, Differential Weight Procedure of the Conditional P.D.F. Approach will take an important role, for it will be used not only for estimating the operating characteristics of correct answers but of any discrete item responses, including the distractors of multiple-choice test items.

Also the content-based observation of informative distractors, which has been described in Chapter 7, will become useful and important. The suggested strategies of writing test items (cf. Section 7.5) can readily be adopted in the construction of itempools as well as in on-line item calibration in the future research.

[VIII.7] Discussion and Conclusions

The above sections have summarized the research accomplishments which will directly contribute to the computerised adaptive testing. Since each accomplishment has been observed and discussed in detail in the previous chapters, this chapter has to be brief.

Efficient computerized adaptive testing is one of the main objectives of the present research. The author has been pleased to introduce these accomplishments that will benefit it from various angles.

References

[1] Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measure-*

ment, 1, 1977, 233-247.

IX Other Findings in the Present Research

There are many other research findings in the present research which have not been reported in the ONR research reports. They concern those topics that are still being pursued, or that will find their places in a more comprehensive framework in the future research.

Among those research findings are those of *winsorization* of the outliers of the maximum likelihood estimates of θ adopted in the process of the Simple Sum Procedure of the Conditional P.D.F. Approach for estimating the operating characteristics of discrete item responses. The results turned out to be fairly successful. We still need further research on this subject, however, before we can evaluate this variation of the Simple Sum Procedure.

Some considerations and observations have also been made concerning possible applications of the theories and methodologies developed so far in the area of latent trait models. They include the latent trait approach to Rorschach diagnosis based upon the Burstein-Loucks scoring system, and the prospect of applying latent trait models and methodologies accommodating both psychological and neurological factors (cf. Chapter 1).

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
Studiestraede 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. Blaiwes
Code N712
Naval Training Systems Center
Orlando, FL 32813-7100

Dr. Bruce Bloxom
Defense Manpower Data Center
99 Pacific St.
Suite 155A
Monterey, CA 93943-3231

Cdt. Arnold Bohrer
Sectie Psychologisch Onderzoek
Rekruterings-En Selectiecentrum
Kwartier Koningen Astrid
Bruijnstraat
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 [illegible]

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
AFHRL/MOEL
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
P. O. Box 16268
Alexandria, VA 22302-0268

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. Crone
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. Hei-Ki Dong
Bell Communications Research
6 Corporate Place
PYA-1K226
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

Dr. Stephen Dunbar
224B Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. James A. Earles
Air Force Human Resources Lab
Brooks AFB, TX 78235

Dr. Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence, KS 66045

Dr. George Englehard, Jr.
Division of Educational Studies
Emory University
210 Fishburne Bldg.
Atlanta, GA 30322

ERIC Facility-Acquisitions
2440 Research Blvd, Suite 550
Rockville, MD 20850-3238

Dr. Benjamin A. Fairbank
Operational Technologies Corp.
5825 Callaghan, Suite 225
San Antonio, TX 78228

Dr. Marshall J. Farr, Consultant
Cognitive & Instructional
Sciences
2520 North Vernon Street
Arlington, VA 22207

Dr. P-A. Federico
Code 51
NPRDC
San Diego, CA 92152-6800

Dr. Leonard Feldt
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard L. Ferguson
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-MRR
The Pentagon
Washington, DC 20310-0300

Prof. Donald Fitzgerald
University of New England
Department of Psychology
Armidale, New South Wales 2351
AUSTRALIA

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Alfred R. Fregly
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons
Illinois State Psychiatric Inst.
Rm 529W
1601 W. Taylor Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Drew Gitomer
Educational Testing Service
Princeton, NJ 08541

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Sherrie Gott
AFHRL/MOMJ
Brooks AFB, TX 78235-5601

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Michael Habon
DORNIER GMBH
P.O. Box 1420
D-7990 Friedrichshafen 1
WEST GERMANY

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Grant Henning
Senior Research Scientist
Division of Measurement
Research and Services
Educational Testing Service
Princeton, NJ 08541

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 63
San Diego, CA 92152-6800

Dr. Thomas M. Hirsch
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Paul W. Holland
Educational Testing Service,
21-T
Rosedale Road
Princeton, NJ 08541

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 92010

Ms. Julia S. Hough
Cambridge University Press
40 West 20th Street
New York, NY 10011

Dr. William Howell
Chief Scientist
AFHRL/CA
Brooks AFB, TX 78235-5601

Dr. Lloyd Humphreys
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Steven Hunka
3-104 Educ. N.
University of Alberta
Edmonton, Alberta
CANADA T6G 2G5

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Robert Jannarone
Elec. and Computer Eng. Dept.
University of South Carolina
Columbia, SC 29208

Dr. Kumar Joag-dev
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign, IL 61820

Dr. Douglas H. Jones
1280 Woodfern Court
Toms River, NJ 08753

Dr. Brian Junker
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Michael Kaplan
Office of Basic Research
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Milton S. Katz
European Science Coordination Office
U.S. Army Research Institute
Box 65
FPO New York 09510-1500

Prof. John A. Keats
Department of Psychology
University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. Jwa-keun Kim
Department of Psychology
Middle Tennessee State University
P.O. Box 522
Murfreesboro, TN 37132

Mr. Soon-Hoon Kim
Computer-based Education Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Dr. William Koch
Box 7246, Meas. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. Richard J. Koubek
Department of Biomedical & Human Factors
139 Engineering & Math Bldg.
Wright State University
Dayton, OH 45435

Dr. Leonard Kroeker
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Dr. Jerry Lehnus
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Mr. Rodney Lim
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

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

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Dr. Richard Luecht
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. George B. Macready
Department of Measurement Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

Dr. Clessen J. Martin
Office of Chief of Naval Operations (OP 13 F)
Navy Annex, Room 2832
Washington, DC 20350

Dr. James R. McBride
The Psychological Corporation
1250 Sixth Avenue
San Diego, CA 92101

Dr. Clarence C. McCormick
HQ, USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago, IL 60064

Mr. Christopher McCusker
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert McKinley
Educational Testing Service
Princeton, NJ 08541

Mr. Alan Mead
c/o Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Timothy Miller
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Dr. William Montague
NPRDC Code 13
San Diego, CA 92152-6800

Ms. Kathleen Moreno
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Headquarters Marine Corps
Code MPI-20
Washington, DC 20380

Dr. Ratna Nandakumar
Educational Studies
Willard Hall, Room 213E
University of Delaware
Newark, DE 19716

Library, NPRDC
Code P201L
San Diego, CA 92152-6800

Librarian
Naval Center for Applied Research
in Artificial Intelligence
Naval Research Laboratory
Code 5510
Washington, DC 20375-5000

Dr. Harold F. O'Neil, Jr.
School of Education - WPH 801
Department of Educational Psychology & Technology
University of Southern California
Los Angeles, CA 90089-0031

Dr. James B. Olsen
WICAT Systems
1875 South State Street
Orem, UT 84058

Office of Naval Research, Code 1142CS
800 N. Quincy Street
Arlington, VA 22217-5000
(6 Copies)

Dr. Judith Orasanu
Basic Research Office
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

Dr. Peter J. Pashley
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Circle, NW
Washington, DC 20036

Dr. James Paulson
Department of Psychology
Portland State University
P.O. Box 751
Portland, OR 97207

Dept. of Administrative Sciences Code 54
Naval Postgraduate School
Monterey, CA 93943-5026

Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Malcolm Ree
AFHRL/MOA
Brooks AFB, TX 78235

Mr. Steve Reiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
310B Austin Peay Bldg.
Knoxville, TN 37916-0900

Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152-6800

Lowell Schoer
Psychological & Quantitative Foundations
College of Education
University of Iowa
Iowa City, IA 52242

Dr. Mary Schratz
905 Orchid Way
Carlsbad, CA 92009

Dr. Dan Segall
Navy Personnel R&D Center
San Diego, CA 92152

Dr. Robin Shealy
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

Dr. Randall Shumaker
Naval Research Laboratory
Code 5510
4555 Overlook Avenue, S.W.
Washington, DC 20375-5000

Dr. Richard E. Snow
School of Education
Stanford University
Stanford, CA 94305

Dr. Richard C. Sorensen
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Judy Spray
ACT
P.O. Box 168
Iowa City, IA 52243

Dr. Martha Stocking
Educational Testing Service
Princeton, NJ 08541

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

Dr. William Stout
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Hariharan Swaminathan
Laboratory of Psychometric and Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

Mr. Brad Sympson
Navy Personnel R&D Center
Code-62
San Diego, CA 92152-6800

Dr. John Tangney
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Kikumi Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Maurice Tatsuoka
220 Education Bldg
1310 S. Sixth St.
Champaign, IL 61820

Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

Mr. Thomas J. Thomas
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Mr. Gary Thomasson
University of Illinois
Educational Psychology
Champaign, IL 61820

Dr. Robert Tsutakawa
University of Missouri
Department of Statistics
222 Math. Sciences Bldg.
Columbia, MO 65211

Dr. Ledyard Tucker
University of Illinois
Department of Psychology
603 E. Daniel Street
Champaign, IL 61820

Dr. David Vale
Assessment Systems Corp.
2233 University Avenue
Suite 440
St. Paul, MN 55114

Dr. Frank L. Vicino
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Howard Wainer
Educational Testing Service
Princeton, NJ 08541

Dr. Michael T. Waller
University of Wisconsin-Milwaukee
Educational Psychology Department
Box 413
Milwaukee, WI 53201

Dr. Ming-Mei Wang
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Thomas A. Warm
FAA Academy AAC934D
P.O. Box 25082
Oklahoma City, OK 73125

Dr. Brian Waters
HumRRO
1100 S. Washington
Alexandria, VA 22314

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman
Box 146
Carmel, CA 93921

5/1/90

Major John Welsh
AFHRL/MOAN
Brooks AFB, TX 78223

Dr. Douglas Wetzel
Code 51
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Rand R. Wilcox
University of Southern California
Department of Psychology
Los Angeles, CA 90089-1061

German Military Representative
ATTN: Wolfgang Wildgrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington, DC 20016

Dr. Bruce Williams
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

Dr. Hilda Wing
Federal Aviation Administration
800 Independence Ave, SW
Washington, DC 20591

Mr. John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. George Wong
Biostatistics Laboratory
Memorial Sloan-Kettering Cancer Center
1275 York Avenue
New York, NY 10021

Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
Code 51
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto
02-T
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Dr. Joseph L. Young
National Science Foundation
Room 320
1800 G Street, N.W.
Washington, DC 20550

Mr. Anthony R. Zara
National Council of State Boards of Nursing, Inc.
625 North Michigan Avenue
Suite 1544
Chicago, IL 60611