

LORD and NOVICK

STATISTICAL  
THEORIES OF  
MENTAL  
TEST  
SCORES

WITH CONTRIBUTIONS BY

A. BIRNBAUM



**STATISTICAL THEORIES  
OF MENTAL TEST SCORES**

The Addison-Wesley Series in  
BEHAVIORAL SCIENCE: QUANTITATIVE METHODS

FREDERICK MOSTELLER, *Consulting Editor*

This series has been undertaken for the publication of works that combine mathematical, statistical, and computational techniques with substantive material. The series will help to integrate these theoretical techniques with disciplines that are concerned with actual problems in the real world. The approach of each book is therefore interdisciplinary, and heavy emphasis falls on the handling of data, on the application of mathematical and statistical devices to concrete issues, and on the development of "cross-departmental" subject matter. Many kinds and many levels of mathematics are used and the range of subjects treated is broad. This wide coverage will benefit all who apply quantitative methods in the area of behavioral science.

FREDERIC M. LORD and MELVIN R. NOVICK  
*Educational Testing Service*

# STATISTICAL THEORIES OF MENTAL TEST SCORES

*with contributions by*  
ALLAN BIRNBAUM, *New York University*

ISBN: 978-1-59311-934-8

Copyright © 2008 IAP - Information Age Publishing Inc.  
All rights reserved. No part of this publication may be reproduced, stored in a  
retrieval system, or transmitted in any form or by any electronic or mechanical means,  
or by photocopying, microfilming, recording or otherwise without written permission  
of the publisher.

Printed in the United States of America

The writing of this book, together with some of the research reported herein, was supported in part by the Logistics and Mathematical Statistics Branch of the Office of Naval Research under contract Nonr-4866(00), NR 042-249, and by the Personnel and Training Branch of the Office of Naval Research under contract Nonr-2752(00), NR 151-201. Reproduction, translation, publication, use, and disposal in whole or in part, by or for the United States Government is permitted.

## PREFACE

More than fifteen years have passed since the publication of Harold Gulliksen's *Theory of Mental Tests*. During that period mental test theory has been developing at an accelerating pace. As a result it has been clear for some time that a new synthesis of the field would be desirable. Professor Gulliksen has recognized this need for many years, and it is partly as a result of his suggestion and encouragement that this book has been written. These years have also seen an increased dependence of the theory on relevant mathematical statistical models, and hence it was apparent to us that any up-to-date, comprehensive treatment of test theory must be based on the statement and development of these models.

Our initial efforts in seeking a new synthesis of the field suggested that a very careful mathematical statement of the relevant models would provide additional insight into many well-studied problems. In preparing this book, consequently, we have tried to exercise more care in this direction than has previously been attempted. We have also tried, however, to avoid technical mathematical discussions that do not seem to contribute to an understanding of test theory. As a result we have been able to keep parts of the book at a more moderate mathematical level than had at one point seemed possible.

A major part of our task has been to restate and refine the work of many people into one integrated development. Our book, however, represents the theoretical orientation and interests of its authors. The relative amounts of space assigned to various topics is due, in part, to this. Published research that is central to our purpose has been integrated into our treatment, and thus we have encompassed much of the test theory literature. Some heretofore unpublished research of our own has also been included. Published research that is relevant but not central to our approach has usually been referred to but not discussed. Some basic papers and many special ones have not been referred to here, either because they do not fit into our approach or because their subject matter is beyond the scope of our undertaking. Thus the inclusion or exclusion of reference to any particular paper should not be interpreted as a comment on the merit of that paper.

During the planning of this book, we recognized that one important piece of research, which at that time was still unpublished, would need to be covered in any comprehensive treatment of mental test theory. This research was Allan Birnbaum's work in latent trait theory, including his logistic response model.

We have been fortunate indeed that Professor Birnbaum has agreed to publish this material for the first time in our book. Professor Birnbaum has had a very free hand in developing his contribution in his own way. However, we have worked closely with him in an endeavor to integrate his material into the general stream of development of our book. The remaining chapters of the book were written by the undersigned, and the responsibility for these chapters, the selection and integration of the work of the contributor, the outline, format, and point of view of the book are jointly ours alone.

Many of the ideas presented here were nurtured in the ongoing Seminar in Mathematical Psychology of the Psychometric Research Group of Educational Testing Service. Members of this Seminar at various times have been Allan Birnbaum, Michael W. Browne, Karl Jöreskog, Walter Kristof, Michael Levine, Frederic M. Lord, William Meredith, Samuel Messick, Roderick McDonald, Melvin R. Novick, Fumiko Samejima, J. Philip Sutcliffe, and Joseph L. Zinnes. Much of our discussion of the relationship between statistical models and the methodological and theoretical problems of psychology has benefited from conversations with Norman Frederiksen, Nathan Kogan, and Samuel Messick. Part III has benefited from a careful reading by Julian Stanley. We are indebted to Professor Louis Guttman for making available to us a copy of his unpublished 1953 manuscript, *The Concepts and Algebra of Reliability and Validity: A Critique*. The manuscript as a whole has been reviewed by Frederick Mosteller and Robert L. Thorndike, from whose suggestions we have profited immensely and to whom we extend our gratitude, but whom we saddle with no responsibility for any errors that may remain.

Preliminary versions of a number of chapters of this book have been used during the past two years at the following universities as a text for courses and seminars in test theory and for related courses.

<i>University</i>	<i>Instructors</i>
University of Chicago	R. Darrell Bock David Wiley Benjamin Wright
Harvard University	John B. Carroll
University College, London	Melvin R. Novick
University of North Carolina	Murray Aitkin
Ontario Institute for Studies in Education	Ross E. Traub
University of Pennsylvania	Melvin R. Novick
Princeton University	Frederic M. Lord Melvin R. Novick
Stanford University	Lee J. Cronbach
University of Tennessee	Edward Cureton

Comments received from those who have used this material, both instructors and students, have been most helpful in preparing later drafts.

Mrs. Dorothy Thayer played a substantial role in checking the manuscript for errors and ambiguities and has had a major responsibility for developing or checking most of the exercises and for preparing many of the tables and charts. Mr. Charles Lewis offered many penetrating comments which have impelled us to sharpen our arguments in many places. Messrs. Carl Frederiksen, Jon Kettenring, Philip Piserchia, and Larry G. Richards, working as summer research assistants, have helped us with early chapter drafts by pointing out errors and ambiguities. The very tedious job of typing the successive drafts of this manuscript has been expertly performed by Mrs. Beatrice Stricklin and Miss Kathleen Rohe, with occasional assistance from Mrs. Mary Evelyn Runyon and Mrs. Eleanor Hibbs. Mrs. Fay F. Richardson assisted with the galley proofs. Mrs. Ann King monitored the final page proofs and provided invaluable editorial assistance during the entire period of this project. Mr. Michael Friendly has carefully checked a number of chapters for typographical errors. Mrs. Sara B. Matlack arranged for and monitored all of the extensive support facilities required for the completion of this project.

We would like to thank all the people at Addison-Wesley who have given us so much fine technical assistance in the completion of this project. Their un-failing sympathy with our many wishes and their constant efforts to comply with these wishes whenever possible, has earned our very deep appreciation.

We are indebted to the Office of Naval Research for extensive support of some of the research reported here. This test theory research has been partly supported by the Personnel and Training Branch since 1952. The preparation of this book was financed, in part, by support received over a three-year period from the Logistics and Mathematical Statistics Branch.

A very large measure of support for this project was provided by Educational Testing Service, to whom we are greatly indebted for making this book possible. In particular, we wish to express our gratitude to William W. Turnbull, Executive Vice-President, and Norman Frederiksen, Director of the Division of Psychological Studies, for creating and maintaining an environment in the Division of Psychological Studies in which basic research can flourish.

*Princeton, New Jersey  
November 1967*

Frederic M. Lord  
Melvin R. Novick



# CONTENTS

## PART 1

### The Foundations of Mental Measurement Theory

---

#### Chapter 1

##### Measurement in Psychology and Education

1.1	The need for a theory of mental testing . . . . .	13
1.2	Psychological theory and its function . . . . .	15
1.3	Measurement as a basis of model construction . . . . .	16
1.4	The place of measurement in psychology . . . . .	19
1.5	Levels of measurement . . . . .	20
1.6	The specification of interval scales . . . . .	21
1.7	Deterministic and probabilistic models . . . . .	23
1.8	The assumptions underlying test theory models . . . . .	24

#### Chapter 2

##### The Construction of True and Error Scores

2.1	Introduction . . . . .	27
2.2	The distribution of measurements on a fixed person . . . . .	29
2.3	True score as an expectation . . . . .	30
2.4	The construction of the error random variable for a fixed person . . . . .	31
2.5	The random selection of persons . . . . .	32
2.6	Construction of the linear model . . . . .	34
2.7	Derivation of the usual assumptions of the classical model . . . . .	36
2.8	What is error? . . . . .	38
2.9	The many concepts of true score . . . . .	39
2.10	Experimental independence . . . . .	44
2.11	Linear experimental independence . . . . .	45
2.12	Replicate measurements . . . . .	46
2.13	Parallel measurements and parallel forms . . . . .	47

**PART 2**  
**The Classical Test Theory Model**

---

**Chapter 3**

**Basic Equations of the Classical Model for Tests of Fixed Length**

3.1	The classical linear model: restatement of assumptions . . . . .	55
3.2	Expectations, variances, and correlations . . . . .	56
3.3	Relationships based on parallel measurements . . . . .	58
3.4	Definitions, interpretations, and applications . . . . .	60
3.5	The validities of a test . . . . .	61
3.6	An alternative statement of the classical model . . . . .	63
3.7	Regression theory for the classical model . . . . .	64
3.8	Errors of measurement, estimation, and prediction . . . . .	66
3.9	Attenuation formulas . . . . .	69
3.10	Elementary models for inferring true change . . . . .	74

**Chapter 4**

**Composite Tests**

4.1	Introduction . . . . .	82
4.2	Composite measurements with two components . . . . .	83
4.3	Composite measurements with $n$ components . . . . .	85
4.4	Coefficient $\alpha$ and the reliability of composite measurements . . . . .	87
4.5	The internal structure of tests . . . . .	95
4.6	Expectations, variances, and covariances of weighted composites . . . . .	96
4.7	The correlation between two composite measurements . . . . .	97

**Chapter 5**

**Basic Equations of the Classical Model  
for Homogeneous Tests of Variable Length**

5.1	Test length as a test parameter . . . . .	103
5.2	The classical model with a continuous test length parameter . . . . .	104
5.3	Statement of the assumptions . . . . .	105
5.4	The true score as the observed score of a person on a test of infinite length . . . . .	108
5.5	The fundamental theorem . . . . .	108
5.6	Expectations and variances . . . . .	109
5.7	Covariances . . . . .	110
5.8	Correlations among observed, true, and error scores . . . . .	111
5.9	Expectations, variances, and correlations of lengthened tests . . . . .	111
5.10	The Spearman-Brown formula . . . . .	112
5.11	The effect of test length on validity . . . . .	114
5.12	Comparing reliabilities and validities of tests of differing lengths . . . . .	118
5.13	The most reliable composite with a specified true score . . . . .	119

5.14	Maximizing the reliability of the composite when component lengths are fixed . . . . .	123
5.15	Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time . . . . .	124

**Chapter 6****Factors Affecting Measurement Precision,  
Estimation, and Prediction**

6.1	Introduction . . . . .	129
6.2	Effect of group heterogeneity on test reliability . . . . .	129
6.3	Speed and power tests . . . . .	131
6.4	Conditions of measurement affecting reliability . . . . .	133
6.5	Experimental problems in correcting for attenuation . . . . .	137
6.6	Accuracy of the Spearman-Brown prophecy formulas . . . . .	139
6.7	Reliability as a generic concept . . . . .	139
6.8	Effect of explicit and incidental selection on test validity: the two-variable case . . . . .	140
6.9	The effect of selection on test validity: the three-variable case . . . . .	144
6.10	The effect of selection on test validity: the general case . . . . .	146
6.11	Accuracy of the selection formulas . . . . .	147

**Chapter 7****Some Estimates of Parameters of the Classical Model**

7.1	Introduction . . . . .	151
7.2	Estimating true score . . . . .	152
7.3	An unbiased estimate of the specific error variance from parallel measurements . . . . .	153
7.4	The use of estimated error variances . . . . .	159
7.5	Specific true-score variance estimated from an analysis of variance components . . . . .	160
7.6	A general formulation of the estimation problem as an analysis of variance components . . . . .	162
7.7	An estimate of an upper bound on the specific error variance from measurements that are not strictly parallel . . . . .	166

**PART 3****Other Weak True-Score Models****Chapter 8****Some Test Theory for Imperfectly Parallel Measurements**

8.1	Defining true score . . . . .	173
8.2	The generic error of measurement . . . . .	176
8.3	The generic error variance . . . . .	177

8.4	Basic properties of generic errors of measurement . . . . .	180
8.5	Generic true-score variance . . . . .	184
8.6	The relation between generic and specific true-score variances . . . . .	185
8.7	Estimating generic parameters describing a single test form . . . . .	187
8.8	Comparisons of estimates of error variance . . . . .	191
8.9	Substantive considerations regarding choice among estimates . . . . .	194

**Chapter 9****Types of Reliability Coefficients and Their Estimation**

9.1	Introduction . . . . .	198
9.2	Estimating the specific reliability coefficient . . . . .	200
9.3	Statistical properties of an estimated variance ratio . . . . .	201
9.4	Specific reliability theory for composite tests . . . . .	203
9.5	Maximum likelihood estimation of reliability for normally distributed scores . . . . .	204
9.6	The frequency distribution of the estimated reliability . . . . .	206
9.7	The generic reliability coefficient . . . . .	208
9.8	Generic reliability for a single test . . . . .	209
9.9	Use and interpretation of reliability coefficients . . . . .	211
9.10	The reliability of ordinal measurements . . . . .	214
9.11	Use of factor loadings as reliability coefficients . . . . .	216
9.12	Estimating reliability without using parallel forms . . . . .	216

**Chapter 10****Some Test Theory for  $\tau$ -Equivalent Measurements,  
Including Estimation of Higher-Order Moments**

10.1	Introduction and definitions . . . . .	224
10.2	An assumption of linear experimental independence . . . . .	225
10.3	Immediate implications . . . . .	226
10.4	Basic theorem for $\tau$ -equivalent measurements . . . . .	227
10.5	Third-order moments . . . . .	228
10.6	Higher-order moments and cumulants . . . . .	230
10.7	Regression of true score on observed score . . . . .	230
10.8	Implications, applications, and limitations . . . . .	232

**Chapter 11****Item Sampling in Test Theory and in Research Design**

11.1	Introduction . . . . .	234
11.2	Matrix sampling . . . . .	236
11.3	Generalized symmetric means . . . . .	238
11.4	First- and second-degree gsm's . . . . .	241
11.5	Estimating true-score moments . . . . .	245
11.6	Estimating the relation of observed score to true score . . . . .	248
11.7	Estimating the relation between scores on parallel test forms . . . . .	248

11.8	Estimating the observed-score statistics for lengthened tests . . . . .	249
11.9	Frequency distribution of errors of measurement for binary items . . . . .	250
11.10	Item sampling as a technique in research design . . . . .	252
11.11	Estimating a mean from a single item sample . . . . .	253
11.12	Estimating a mean by multiple matrix sampling . . . . .	255
11.13	Estimating group mean differences . . . . .	257
11.14	Estimating observed-score variances by item sampling . . . . .	259

---

**PART 4**  
**Validity and Test Construction Theory**

---

**Chapter 12**

**Validity**

12.1	Introduction . . . . .	261
12.2	Regression and prediction . . . . .	262
12.3	Linear regression functions . . . . .	264
12.4	Multiple and partial correlation . . . . .	265
12.5	Partial and multiple correlation and regression in $n$ variables . . . . .	267
12.6	The screening of predictor variables . . . . .	269
12.7	Suppressor variables, moderator variables, and differential predictability . . . . .	271
12.8	Incremental validity . . . . .	273
12.9	Validity and the selection ratio . . . . .	275
12.10	Some remarks on the explication of the concept of validity as a correlation coefficient . . . . .	277
12.11	Construct validity . . . . .	278

**Chapter 13**

**The Selection of Predictor Variables**

13.1	Introduction . . . . .	284
13.2	Some sampling problems . . . . .	284
13.3	Formal procedures for selecting predictor variables . . . . .	288
13.4	Prediction in future samples . . . . .	289
13.5	The effect of relative test lengths on reliability and validity: the multiple predictor case . . . . .	293
13.6	The determination of relative test lengths to maximize the multiple correlation . . . . .	295

**Chapter 14**

**Measurement Procedures and Item-Scoring Formulas**

14.1	Introduction . . . . .	302
14.2	Guessing and omitting . . . . .	303

14.3	A simple formula score . . . . .	305
14.4	Properties of the simple formula score . . . . .	307
14.5	A simple regression model for scoring items . . . . .	310
14.6	The regression method with a simple model that assumes partial knowledge . . . . .	312
14.7	Other item-scoring formulas . . . . .	313
14.8	An evaluation of partial knowledge . . . . .	314
14.9	Methods for discriminating levels of partial knowledge concerning a test item . . . . .	315
14.10	Assumptions underlying the personal probability approach to item scoring . . . . .	319
14.11	Reproducing scoring systems . . . . .	321

## **Chapter 15**

### **Item Parameters and Test Construction**

15.1	Introduction . . . . .	327
15.2	Item difficulty . . . . .	328
15.3	Item discriminating power . . . . .	329
15.4	Item validity . . . . .	332
15.5	Product moment correlations for dichotomous items . . . . .	335
15.6	Biserial correlation . . . . .	337
15.7	Comparison of biserial and point biserial coefficients . . . . .	340
15.8	Tetrachoric correlation . . . . .	345
15.9	A comparison of tetrachoric and phi coefficients . . . . .	346
15.10	Considerations in the choice of test construction techniques . . . . .	350
15.11	Formula scoring and corrections for chance success . . . . .	352
15.12	Invariant item parameters . . . . .	353

## **Chapter 16**

### **Latent Traits and Item Characteristic Functions**

16.1	Introduction . . . . .	358
16.2	Latent variables . . . . .	359
16.3	Local independence . . . . .	360
16.4	Item-test regression . . . . .	363
16.5	The normal ogive model . . . . .	365
16.6	Conditions leading to the normal ogive model . . . . .	370
16.7	Correlation matrix with one common factor . . . . .	371
16.8	A sufficient condition for normal ogive item characteristic curves . . . . .	374
16.9	Normal ogive parameters: item difficulty . . . . .	376
16.10	Normal ogive parameters: item discriminating power . . . . .	377
16.11	Practical use of normal ogive item parameters . . . . .	379
16.12	Conditional distribution of test scores . . . . .	384
16.13	A relation of latent trait to true score . . . . .	386
16.14	Typical distortions in mental measurement . . . . .	387

**PART 5****Some Latent Trait Models  
and Their Use in Inferring an Examinee's Ability***(Contributed by Allan Birnbaum)***Chapter 17****Some Latent Trait Models**

17.1	Introduction . . . . .	397
17.2	The logistic test model . . . . .	399
17.3	Other models . . . . .	402
17.4	The test as a measuring instrument: examples of classification and estimation of ability levels by use of test scores . . . . .	405
17.5	Information structure of a test and transformations of scale of scores . . . . .	410
17.6	Transformations of scales of ability . . . . .	411
17.7	Calculations of distributions of test scores . . . . .	414
17.8	Quantal response models in general . . . . .	420
17.9	Estimation of item parameters . . . . .	420
17.10	Validity of test models . . . . .	422

**Chapter 18****Test Scores, Sufficient Statistics,  
and the Information Structures of Tests**

18.1	Sufficient statistics: definition and interpretation . . . . .	425
18.2	Conditions for sufficiency of a statistic . . . . .	428
18.3	Test scores and sufficient statistics . . . . .	429
18.4	Sufficiency and the logistic test model . . . . .	431
18.5	Sufficiency and the information structures of tests . . . . .	434

**Chapter 19****Classification by Ability Levels**

19.1	Classification rules for distinguishing two levels of ability . . . . .	436
19.2	Two-point classification problems . . . . .	437
19.3	Locally best weights and classification rules . . . . .	442
19.4	More general classification rules, composite scores, and statistical efficiency in general . . . . .	444
19.5	Quantitative appraisal and efficient design of classification rules . . . . .	446

**Chapter 20****Estimation of an Ability**

20.1	Introduction . . . . .	453
20.2	Some algebra of information functions . . . . .	453
20.3	More general methods of estimation: maximum likelihood . . . . .	455
20.4	The information functions of various test items . . . . .	460



# INTRODUCTION

Our primary goal in this book is to sharpen the skill, sophistication, and intuition of the reader in the interpretation of mental test data, and in the construction and use of mental tests both as instruments of psychological theory and as tools in the practical problems of selection, evaluation, and guidance. We seek to do this by exposing the reader to some psychologically meaningful statistical theories of mental test scores.

Although this book is organized in terms of test-score theories and models, the practical applications and limitations of each model studied receive substantial emphasis, and these discussions are presented in as nontechnical a manner as we have found possible. Since this book catalogues a host of test theory models and formulas, it may serve as a reference handbook. Also, for a limited group of specialists, this book aims to provide a more rigorous foundation for further theoretical research than has heretofore been available.

One aim of this book is to present statements of the assumptions, together with derivations of the implications, of a selected group of statistical models that the authors believe to be useful as guides in the practices of test construction and utilization. With few exceptions we have given a complete proof for each major result presented in the book. In many cases these proofs are simpler, more complete, and more illuminating than those originally offered. When we have omitted proofs or parts of proofs, we have generally provided a reference containing the omitted argument. We have left some proofs as exercises for the reader, but only when the general method of proof has already been demonstrated. At times we have proved only special cases of more generally stated theorems, when the general proof affords no additional insight into the problem and yet is substantially more complex mathematically.

We have attempted to present the selected test theory models in a sequence that emphasizes certain groupings, these groupings being determined by the nature of the assumptions made in the models. This ordering leads to the explication of a number of test-score theories. We indicate interrelationships of these theories, and further integrate them into a general theory of latent traits. This integration is the underlying theme of the book.

An important task undertaken in this book is to provide explicit syntactic (i.e., mathematical) definitions for concepts having semantic (i.e., real-world) significance. Foremost among these is the concept of the true score, which

previously has usually been defined syntactically and interpreted semantically as the observed score that a person would obtain on an infinitely long test. In this book the true score is, for most purposes, defined syntactically as an expected observed score. It is then shown, by means of the law of large numbers, that the previous test theoretic definition is a valid semantic interpretation. The authors feel that many of the controversies that have continued to plague test theory disappear when rigorous syntactic and semantic definitions for such basic test theoretic concepts as true score and error are explicated and differentiated.

Since our primary interest is the scientific rather than the technological applications of test theory, the statistical methods used in this book fall primarily within that domain sometimes called informative inference. Some classical methods of point estimation and confidence interval estimation as methods of informative inference are used extensively. Those of Bayesian persuasion know that these estimates are often good approximations to similar estimates derived from Bayesian analyses.

We have not presented a survey of decision theoretic approaches to test theory problems. Although decision theoretic models are now extremely useful conceptually and we hope will ultimately be valuable in applications to testing problems, we do not feel that the available models are in general sufficiently flexible or realistic for immediate formal application. The injudicious use of overly formal models and the abdication of decision-making responsibility by the subject matter specialist that are encouraged by current decision theoretic formulations should, we feel, be discouraged, particularly in a scientific context. We share the judgment of Cronbach and Gleser (1965, vii) that currently "decision theory is more important as a point of view than as a source of formal mathematical techniques for developing and applying tests". The interested reader is referred to *Psychological Tests and Personnel Decisions* by Cronbach and Gleser (1965) and to *Studies in Item Analysis and Prediction* edited by Herbert Solomon (1961). On the other hand, we view the work of Birnbaum contained in Part V as an attempt to bridge the gap between inferential and decision theoretic formulations. Also we show in a number of places that certain standard inference procedures are better understood from a decision theoretic point of view.

A major problem in educational testing that militates against the formal use of decision theory is the typical lack of a simple measurable criterion on which to base an analysis. For example, if we are to select students for medical school, we desire to select those who, after the prescribed training, will be the "best doctors". But unfortunately we have no clear-cut measure of what constitutes a good doctor, just as we have no clear-cut measure of what constitutes a good teacher. Too often the criterion we use has availability as its major virtue. This criterion problem permeates all prediction problems of educational testing. For example, see Kelly (1964).

Another problem involved in using decision theory is the great difficulty in giving value measures to various levels of any chosen criterion. Even if we as-

sume that the future value of a medical school graduate is related monotonically to his score on the National Board of Medical Examiners Test, it is difficult to justify the assumption that this relationship is linear, or quadratic, or of decreasing marginal utility. Yet current formulations of statistical decision theory require that we specify the nature of this relationship rather precisely.

There is at least one other serious problem limiting the applicability of certain well-developed statistical methods. With only a few very specialized exceptions, tests are designed for use under a wide range of differing conditions. No test publisher can validate each of his tests under every condition of possible application. (Each test user should, of course, validate locally whatever published tests he intends to employ on a continuing basis.) Hence, a purely predictive approach to test theory has not proved entirely satisfactory.

Both test publishers and psychologists have recognized the difficulties inherent in formal decision theoretic and predictive approaches, and therefore have retained, expanded, and refined the approach to test development that formed the basis for the first mental tests. This approach is based on a theoretical system that incorporates the concepts of aptitudes and abilities, which are presumed to affect a person's performance on mental tests and also his performance on various criteria of interest. In modern psychological theory we think in terms of many kinds of aptitudes and abilities. Test development technology is concerned with the development of tests that measure the aptitudes and abilities of interest with the least possible error. Latent trait theory is concerned with the identification of those traits (aptitudes, abilities, interests, and personality characteristics) that have wide relevance to human behavior, and with the use of measures of these traits to explain and predict human behavior. Although the major emphasis in this book is on theories of measurement of and inference about latent traits, we shall survey those aspects of prediction theory that have proved of obvious direct relevance to testing applications.

Since this book is concerned primarily with test theory, its coverage of applications to specialized problems of personnel selection is incomplete. Horst's (1955) monograph, "A technique for the development of a differential prediction battery", and Thorndike's (1949) *Personnel Selection* may be used as supplements to our abbreviated discussions in this area. The American Council on Education text, *Educational Measurement*, edited in its second edition by Thorndike (in preparation), may be read for its extensive discussion of the practical and theoretical problems of educational measurement. The American Psychological Association publication, *Standards for Educational and Psychological Tests and Manuals* (1966), is a valuable guide for anyone concerned with mental testing.

In a limited sense our book is a text or training book, and indeed the presentation here is more expository than would be necessary or desirable in a purely formal theoretical treatise. Early chapters of the book tend to be somewhat less demanding, technically and conceptually, than later chapters. Also the reader is generally given the opportunity to master concrete material before

he is confronted with abstract concepts. Many fine points of theory are reserved for the exercises, in which many of the numerical examples will also be found. While this book does not present itself primarily as a textbook in test theory, it has been shown to be possible to fashion various test theory courses around parts of it. It may also serve as a supplementary text for a number of courses (in factor analysis, latent structure theory, etc.) in psychometric theory.

The reader with limited formal mathematical or statistical training will find much useful material here because we have attempted to provide a relatively nontechnical, if quite brief, exposition in parallel with each formal presentation. With the exception of a few sections that we indicate can be omitted without loss of continuity, Chapters 1 through 6 and 12 through 15 require little more than a facility with the algebra of expectations. These chapters, together with Chapter 7, may conveniently be used as the core of a one-semester test theory course. Chapters 8 through 11, and to a lesser extent Chapter 7, will be read more easily by those having some prior exposure to model II analysis of variance, although in fact these chapters are essentially self-contained. Our presumption is that readers wishing to study the remaining chapters of the book will have moderate competence in the differential and integral calculus and a familiarity with the language, basic results, and basic mechanics of mathematical statistics. Except occasionally in those sections that have been labeled as being of an advanced nature and that can be omitted without loss of continuity, whenever we have introduced an advanced mathematical-statistical concept (such as minimal sufficiency) we have given a complete development of it.

Up to a certain point, the stronger the reader's mathematical and statistical preparation, the more he can hope to gain from this book and the more easily he will be able to read the latter parts of it; though we would emphasize that the completion of a sequence of *professional* courses in mathematical statistics is nowhere required. A general familiarity with mathematical statistics at the level of Freeman (1963), Mood and Graybill (1963), or Hoel (1954) suffices for almost all the text. In a few chapters some reference is made to Lindgren (1962) and to Kendall and Stuart (1958, 1961). Although we assume familiarity with these texts or their equivalents, we do not assume that the reader has a professional competence in this material. In practice this means that we have given more detailed derivations than we should have done if we had written the book for professional statisticians.

All this is not to say that all the text will be easy reading for anyone with this mathematical and statistical background. Many sections and a few chapters, which are clearly labeled as being primarily of interest to specialists, will be found much more difficult because of the complexity of the material covered, even though we have used no advanced mathematical or statistical methods.

The psychological prerequisites are somewhat less definable. A systems course in psychology, a course in scientific methodology, a course in applied experimental design, and a course in the applications of psychological tests would all be helpful, but surely not all are essential. A general familiarity with

applied statistics at the level of Hays (1963) and Winer (1962) would also be helpful. For those reading this text without benefit of accompanying motivating lectures, some prior or concurrent exposure to the practical problems of testing would be helpful. If the reader lacks such exposure, he is advised to examine Cronbach's (1960) informative work, *The Essentials of Psychological Testing*, and the collection of 74 papers, *Problems in Human Assessment*, edited by Jackson and Messick (1967).

## References

- American Psychological Association, *Standards for educational and psychological tests and manuals*. Washington, D.C.: American Psychological Association, 1966.
- CRONBACH, L. J., *The essentials of psychological testing*. New York: Harper, 1960.
- CRONBACH, L. J., and GOLDINE C. GLESER, *Psychological tests and personnel decisions*, 2nd ed. Urbana: University of Illinois Press, 1965.
- FREEMAN, H., *Introduction to statistical inference*. Reading, Mass.: Addison-Wesley, 1963.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.
- HALPERIN, M., H. O. HARTLEY, and P. G. HOEL, Recommended standards for statistical symbols and notations. *American Statistician*, 1965, **19**, 12-14.
- HAYS, W. L., *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- HOEL, P. G., *Introduction to mathematical statistics*. New York: Wiley, 1954.
- HORST, P., A technique for the development of a differential prediction battery. *Psychological Monograph*, 1955, **69**, No. 5 (Whole No. 390).
- JACKSON, D. N., and S. MESSICK (Eds.), *Problems in human assessment*. New York: McGraw-Hill, 1967.
- KELLY, E. L., Alternate criteria in medical education and their correlates. In *Proceedings of the 1963 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1964, pp. 64-85. Reprinted in Anne Anastasi (Ed.), *Testing problems in perspective*. Washington, D.C.: American Council on Education, 1966.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. New York: Hafner. Vol. I, 1958; Vol. II, 1961.
- LINDGREN, B. W., *Statistical theory*. New York: Macmillan, 1962.
- MOOD, A. M., and F. A. GRAYBILL, *Introduction to the theory of statistics*. New York: McGraw-Hill, 1963.
- SOLOMON, H. (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961.
- THORNDIKE, R. L., *Personnel selection*. New York: Wiley, 1949.
- THORNDIKE, R. L. (Ed.), *Educational measurement*. New York: American Council on Education, in preparation.
- WINER, B. J., *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

## NOTATION

The notation used in this book conforms in most major respects with the recommendations of the *Committee of Presidents of Statistical Societies* (Halperin, Hartley, and Hoel, 1965) as published in the *American Statistician*, June 1965, although some minor modifications have been introduced. The major features of our notational system are as follows:

1. Capital letters are used to denote random variables. Exceptions are made for some multivariate problems.
2. Lower-case letters are used to denote sample observations or nonrandom variables.
3. Lower-case letters are used to denote probability density functions, for example,  $f(x)$ . The corresponding cumulative probability functions are denoted by upper-case letters, for example,  $F(x)$ .
4. With few exceptions, Greek letters are used to denote population parameters throughout.
5. To indicate a sample estimator of a population parameter and its values (the estimates), the parameter symbol is modified by placing a caret (^) over it.
6. For a random variable  $X$ , the mean value of  $X$  is denoted by  $\mathcal{E}(X)$ , or  $\mathcal{E}X$ , or  $\mu_X$ , with subscript or parenthetic modifiers as necessary, for example,

$$\mu_1 = \mu(X_1) = \mathcal{E}(X_1).$$

7. The  $k$ th moment of a random variable  $X$  is denoted by  $\mathcal{E}(X^k)$  or  $\mathcal{E}X^k$ , possibly with appropriate parenthetic or subscript modifiers.
8. The arithmetic mean of a set of numbers, say  $x_1, \dots, x_n$ , is denoted by  $\bar{x}$ , that is,  $\bar{x} = \sum x_i/n$ .
9. Dot subscripts are used to indicate sample means; for example,

$$x_{\cdot} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \quad x_{g\cdot} = \frac{1}{N} \sum_{a=1}^N x_{ga},$$
$$x_{\cdot a} = \frac{1}{n} \sum_{g=1}^n x_{ga}, \quad x_{\cdot\cdot} = \frac{1}{nN} \sum_{g=1}^n \sum_{a=1}^N x_{ga}.$$

Plus subscripts (+), for example,  $x_{g+}$ , are used to denote corresponding sums.

10. For a random variable  $X$ , the variance of  $X$  is denoted by  $\sigma^2(X)$ ,  $\sigma_X^2$ , or  $\text{Var}(X)$ .
11. For random variables  $X_1, X_2$ , the covariance is denoted by  $\sigma(X_1, X_2)$ ,  $\sigma_{12}$ , or  $\text{Cov}(X_1, X_2)$ . If a notation such as  $X, Y$  is used for the random variables, the symbol  $\sigma_{XY}$  is used. Correlation coefficients are similarly denoted, using  $\rho$  instead of  $\sigma$ .
12. Upper-case boldface letters, for example,  $\mathbf{A}$ , are used to denote matrices. The symbol  $\|a_{ij}\|$  is used to denote a matrix  $\mathbf{A}$  with  $i, j$ th element  $a_{ij}$ .
13. Lower-case boldface letters, for example,  $\mathbf{a}$ , are used to denote column vectors. The symbol  $\{a_i\}$  is used to denote a column vector with  $i$ th element  $a_i$ . Lower-case boldface letters are used with prime superscripts to denote row vectors, for example,  $\mathbf{a}'$ .
14. The matrix of variances and covariances of random variables  $X_1, \dots, X_n$  is denoted by  $\Sigma$ , that is,

$$\Sigma = \|\sigma_{ij}\|.$$

15. The coefficient of multiple correlation between a random variable  $X_0$  and, say,  $n$  other random variables  $X_1, \dots, X_n$  is denoted by  $\rho_{0\cdot 123\dots n}$ .
16. The coefficients of a linear regression function are denoted by the letter  $\beta$ , with appropriate subscripts, and the constant is denoted by  $\alpha$ . Thus the linear regression of  $X_0$  on  $k$  independent variables  $x_1, x_2, \dots, x_k$  may be written as

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k;$$

more explicitly, the  $i$ th weight may be written as

$$\beta_{0i\cdot 12\dots(i-1)(i+1)\dots n}.$$

17. The coefficient of partial correlation between random variables  $X_0$  and  $X_1$  that takes into account the joint distribution of the random variables  $X_2, \dots, X_n$  is denoted by  $\rho_{01\cdot 23\dots n}$ .
18. For a set of  $n$  numbers  $x_1, \dots, x_n$ , the symbol  $s^2$  is defined by

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

19. For a set of  $n$  number pairs  $[(x_{11}, x_{21}), \dots, (x_{1n}, x_{2n})]$ , the symbol  $s_{12}$  is defined by

$$s_{12} = \frac{1}{n} \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2).$$

Similarly

$$r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}.$$

The following is a brief glossary of important symbols used in more than one or two sections.

$a, b$	indexing subscripts for persons
$g, h, i, j$	indexing subscripts for tests (items)
$X_{ga}$	the observed-score random variable for a specified test (item) $g$ and specified person $a$
$\tau_{ga}$	the true score (expected score) of person $a$ on test (item) $g$ ; also an observation of the random variable $T_{ga}$
$E_{ga} = X_{ga} - \tau_{ga}$	the error-score random variable for a specified test (item) $g$ and specified person $a$
$X_{g*}, Y_{g*}$	the observed-score random variable for a specified test (item) $g$ and a randomly selected person
$T_{g*}$	the true-score random variable (over people) for a specified test (item) $g$ ; read “tau-sub-g-star”.
$E_{g*} = X_{g*} - T_{g*}$	the error-score random variable for a specified test (item) $g$
$x_{ga}$	an observation of the random variable $X_{ga}$ or $X_{g*}$
$e_{ga}$	an observation of the random variable $E_{ga}$ or $E_{g*}$
$X, X'$	used jointly to denote parallel tests
$\rho_{XT}^2 = \rho_{XX'}$	the reliability of test $X$ (Section 3.4)
$n$	total number of tests (items)
$N$	total number of persons
$\alpha$	coefficient alpha (Section 4.4)
$T$	total testing time parameter (Section 5.13)
$\xi_a = \mathcal{E}_g \mathcal{E}_k Y_{gak}$ = $\mathcal{E}_g T_{ga}$	generic true score (7.6.1); also used for other kinds of true scores, where specified
$\pi_g = \mathcal{E}_a \mathcal{E}_k Y_{gak}$ = $\mathcal{E}_a T_{ga}$	“item difficulty” = mean item score (7.6.3); for binary items, $\pi_g$ = proportion of correct responses
$\xi_a = n\xi_a$	
$p_g = \frac{1}{N} \sum_{a=1}^N y_{ga}$	sample item difficulty
$q_g = 1 - p_g$	for binary items
$\alpha_{ga} = T_{ga} - \xi_a - \pi_g + \mu_\zeta$	tests-by-persons interaction (7.6.4)
$\tilde{\sigma}_e^2$	generous estimator of specific error variance (7.7.1)
$\epsilon_{ga} = y_{ga} - \xi_a$	generic error variance (8.2.1); $\epsilon_{ga}$ is also used for other kinds of error of measurement, where specified

$U_{ga} = 0 \text{ or } 1$	a binary random variable assuming values $u_{ga}$
$\hat{\rho}_{YT}^2, \hat{\rho}_{XX'}^2,$	
$\tilde{\rho}_{YT}^2, \hat{\rho}_{Y\zeta}^2, \hat{\rho}_{g\zeta}^2$	estimated reliability coefficients (see Table 9.9.1)
$n^{(r)} = n(n - 1) \cdots (n - r + 1) = n!/(n - r)!$	
$z_a = \frac{1}{n} \sum_{g=1}^n u_{ga}$	proportion of items answered correctly by examinee $a$ on an $n$ -item test
$\rho_{gh}, r_{gh}$	product moment correlation between scores on items $g$ and $h$ (Section 15.3)
$\rho_{gX}, r_{gx}$	product moment correlation between item score $Y_{ga}$ and total test score $X_a = \sum_g Y_{ga}$ (Section 15.3)
$\rho_{gv}, r_{gv}$	item validity coefficient, that is, the product moment correlation between item $g$ and criterion $v$ (Section 15.4)
$\Phi(x)$	the cumulative normal distribution function
$\varphi(t)$	the normal density function
$\Psi(x)$	the cumulative logistic distribution function
$\psi(x)$	the logistic probability density function
$\bar{\Phi}(\gamma) = 1 - \Phi(\gamma)$	
$Y'_g$	a continuous latent variable underlying item $g$ (Section 15.6)
$\rho'_{gv}, r'_{gv}$	biserial correlation between item $g$ and criterion $v$ (15.6.6)
$\rho'_{ij}, r'_{ij}$	tetrachoric correlation between items $i$ and $j$ (Section 15.7)
$\pi_{ij}, p_{ij}$	proportion of examinees answering both items $i$ and $j$ correctly
$P_g = P_g(\boldsymbol{\theta}) = \text{Prob}(U_{ga} = 1   \boldsymbol{\theta}),$	the conditional probability of a correct answer to binary item $g$ for a specified value of $\boldsymbol{\theta}$ (16.3.4)
$L_g = L_g(\theta) = a_g(\theta - b_g),$	a linear function of $\theta$ (16.5.2)
$\gamma_i$	a cutting point for variable $Y'_i$ (Section 15.6)
$h_i$	an estimate of $\gamma_i$ (Section 15.6)
$\mathbf{v}$	the vector of item scores for items $1, 2, \dots, n$ (Section 16.8)
$a_g$	a measure of item discriminating power (Section 16.10)
$b_g$	a measure of item difficulty (Section 16.9)

The following special symbols are used in various parts of the text.

<i>Symbol</i>	<i>Meaning</i>
=	is equal to; for example, $x^2 = 4$ . This symbol denotes conditional equality; that is, for <i>some</i> value(s) of the argument ( $x$ ) equality holds.
$\equiv$	is identically equal to; for example, $(x^2 - 1) \equiv (x - 1)(x + 1)$ . Or, is defined as; for example, $\psi(x) \equiv e^x/(1 + e^x)$ . This symbol denotes identity; that is, for <i>all</i> values of the argument ( $x$ ) equality holds. (Used only when there is reason to emphasize that the statement is not a conditional equality.)
$\square$	The proof is now complete.
$\doteq$	is approximately equal to
$\propto$	is proportional to

**Part 1**

**THE FOUNDATIONS**

**OF MENTAL MEASUREMENT THEORY**



## CHAPTER 1

# MEASUREMENT IN PSYCHOLOGY AND EDUCATION

### 1.1 The Need for a Theory of Mental Testing

An instructor who has written a final course examination could probably write another very different from the first and yet equally satisfactory. If a person were to take both examinations, it is not likely that he would obtain the same score on both and he might obtain substantially different scores. In fact, an ordering of a group of students based on an administration of the first would not ordinarily agree with an ordering based on an administration of the other. This variability draws attention to a serious problem: These differences in scores and orderings represent the failure of the measuring instrument to do what we wish it to do. Since we wish to measure the "ability" of the examinee, two perfect measuring instruments should yield the same score for a particular examinee so long as his ability does not change. Since the differences between the scores on the two administrations are irrelevant to our main purpose, they may be said to have been caused by "errors of measurement".

One reason we need to have a theory of mental testing is that mental test scores contain sizable errors of measurement. Another reason is that the abilities or traits that psychologists wish to study are usually not directly measurable; rather they must be studied indirectly, through measurements of other quantities. We cannot directly measure a person's mathematical ability; we can only measure his performance on a number of mathematical test items. Indeed these abilities are in general given mathematical definition only in terms of specified measurements.

There is a classical theory of errors appropriate to measurement in the physical sciences (e.g., see Baird, 1962; Parratt, 1961; Topping, 1955). However, we may note two important differences between measurement problems in the physical and the behavioral sciences. In the first place, in the physical sciences we can usually repeat the same measurement a number of times. In mental testing we can perhaps repeat a measurement once or twice, but if we attempt further repetitions, the examinee's responses change substantially because of fatigue or practice effects. The second important difference is that in the physical sciences it is usual to measure and make inferences about one object (or event) at a time. Perhaps the physicist wants to get an accurate determination of the atomic weight of oxygen. In mental testing, on the other hand,

the psychologist often wants to test a whole group of individuals at one time and to make inferences about them individually and in relation to the group. The logical and statistical problems in making inferences simultaneously about all individuals in a group introduce many complexities.

The economist faces an even more extreme situation. He has a "measure", say, of the national income of the United States in a particular year. He considers that this "measure" is fallible in that it contains errors of measurement; yet unlike the psychologist, he cannot repeat the measurement. Nevertheless the economist (e.g., Reiersøl, 1950) manages to use a theory involving errors and to obtain estimates of the magnitudes of these errors and of the true values that they obscure. The situation in mental test theory lies between that in economics and that in the physical sciences.

Another area in which theory of errors is employed is the analysis of voting behavior (e.g., see Campbell *et al.*, 1966) in which fluctuations from the "normal vote" prevent the direct measurement of that construct. Similar methods are used in genetics (e.g., see Kempthorne, 1957, Chapter 23). For example, the supposed "genetic composition" of a herd of cattle cannot be measured directly, but the effects of this composition can be studied through controlled breeding. An area that has many problems almost identical with some problems in mental test theory is that of statistical astronomy (e.g., see Trumpler and Weaver, 1953). Although the problems and methods differ widely in these fields, they all possess two common characteristics: Each postulates constructs that are not directly measurable and each observes phenomena that manifest these latent constructs but that also exhibit fluctuation due to other factors, which are sometimes called errors.

One function of any measurement theory employing hypothetical constructs and measurement deviations from these hypothetical values is the explanation of "paradoxical events". The following exaggerated situation illustrates a paradox that can be clarified by a knowledge of test theory.

The statistics majors in a certain university are assigned a score on a comprehensive final examination in statistics. These final statistics scores are compared with quantitative aptitude scores and verbal aptitude scores obtained prior to admission to the university. It is found that in the group studied, the correlation between the final statistics score and the verbal aptitude score is slightly higher than the correlation between the final statistics score and the quantitative aptitude score. This apparently shocking situation can be understood and explained in test theoretic terms.

A less extreme paradox appears whenever a new test is recommended as a replacement for a test currently in use. If correlations between predictor and criteria are computed for both new and current tests when the sample has already been selected on the basis of the current test, then the current test is typically shown to great and unwarranted disadvantage. Test theory suggests to us that, typically, if one test is used to define the comparison group, then the correlation between that test and a criterion will be substantially lowered in that

group. In fact, in some situations encountered in practice, the new test will *appear* to be better than the old when it is not. We shall return to these examples and explain these paradoxes in detail in Chapter 6.

## 1.2 Psychological Theory and Its Function

As a science, psychology is organized and unified through psychological theories. A psychological theory is a system of statements concerning a set of concepts, which serves to describe, explain, and predict some limited aspects of the behavioral domain. Examples of such concepts are achievement, anxiety, ability, and attitude. Usually a well-developed theory contains one or more formal models which give concrete structure to the general concepts of the theory. These models may be viewed as explications of portions of the general theory. Such models, in turn, are connected systematically with directly observable phenomena. The function of such models is to permit the logical deduction of general and specific relationships that have not been empirically demonstrated but that may be demonstrable. No scientific theory is now thought to depict absolute truth, but the associated models are taken as more or less accurate and useful in various contexts; the Newtonian theory of mechanics, for example, is useful for making some kinds of predictions but not others.

The elements of a psychological theory, the concepts or constructs, are related to each other through statements which we shall call *syntactic definitions*. These definitions are expressed as formal logical (mathematical) relations. Some of these constructs are related to observable behavior through definitions which we shall call *semantic definitions* (see Torgerson, 1958, who uses the corresponding terms *constitutive* and *epistemic*). These semantic definitions form rules of correspondence between the theoretical constructs and the behavioral domain. Theoretical constructs, whether they are stated in verbal or in verbal-mathematical form, are abstractions from aspects of nature. The syntactic definition of these constructs permits the development of a more or less precisely formulated logical system. This logical system permits the deduction of the properties of the constructs of the system. These constructs may then be interpreted semantically to explain past events or predict future events in the behavioral domain.

In psychology, according to Torgerson (1958):

The concepts of theoretical interest tend to lack empirical meaning, whereas the corresponding concepts with precise empirical meaning often lack theoretical import. One of the great problems in the development of a science is the discovery or invention of constructs that have . . . both.

The more "theoretical" constructs are often not far removed from simple common-sense or prescientific conceptions. Though they have a great deal of common-sense meaning attached to them, the meaning is not specified precisely. The terms are thus somewhat vague, and more often than not are complex. Before a satisfactory state of affairs is reached, it is necessary somehow to transform these inexact, complex concepts into exact ones which can be specified precisely. This is what Carnap (1950)

calls the task of *explication*. Though this task seems common to all sciences, it is particularly acute in those disciplines that are in their initial stages of development. It is especially true at the present time in the social and behavioral sciences, where an immense amount of time has been devoted to construction of complex and elaborate theoretical superstructures based on unexplicated, inexact constructs.

When we devise a rule of correspondence for relating one of these prescientific concepts to observable data, we are in fact carrying out an explication of the original concept. Our purpose is to replace the original concept with one that is defined more precisely . . . (p.8)\*

The distinction between semantic and syntactic definitions is one which we shall emphasize repeatedly and discuss in some further detail in the following section.

### 1.3 Measurement as a Basis of Model Construction

Measurement, in psychology and elsewhere in science, begins with a procedure for identifying elements of the real world with the elements or constructs of an abstract logical system (a model) through the precise semantic definition of the basic elements of the theory. As a simple example, the concept of a person's height is one which is easily defined through the following measurement procedure. To specify this measurement we must do three things: First we must identify the object being measured, the person or experimental unit. Then we must identify the property or behavior being directly measured, in this case the distance between the sole of his foot and the top of his head when he is standing up straight. Finally we must identify the numerical assignment rule by which we assign a number to this property of the unit being measured. In this example, the procedure may be to use, in the customary way, a tape measure calibrated linearly in inches. On a multiple choice test, the behavior being directly measured is the examinee's responses to the items and the numerical assignment rule is often stipulated as the number of "correct answers". Measurement, to some extent, has always been a part of psychology. In recent years, however, psychology has seen an increased use of measurement, and indeed a *systematization* of measurement in the form of psychological models which are primarily mathematical in form.

Often the identification of the elements of the real world with the elements of the mathematical system can be done in a very natural way as in the above example. The following is a second, less obvious example of measurement. Suppose that we flip a coin which may come up heads or tails (assuming that the unlikely event of its coming to rest on its edge does not occur) and that we define two associated events,  $H$  and  $T$ . These events are mutually exclusive and exhaustive in that every possible realization may be associated with one and only one of the two subsets.

---

\* From Torgerson, W. S., *Theory and methods of scaling*, Copyright, 1958. John Wiley & Sons, Inc. Used by permission.

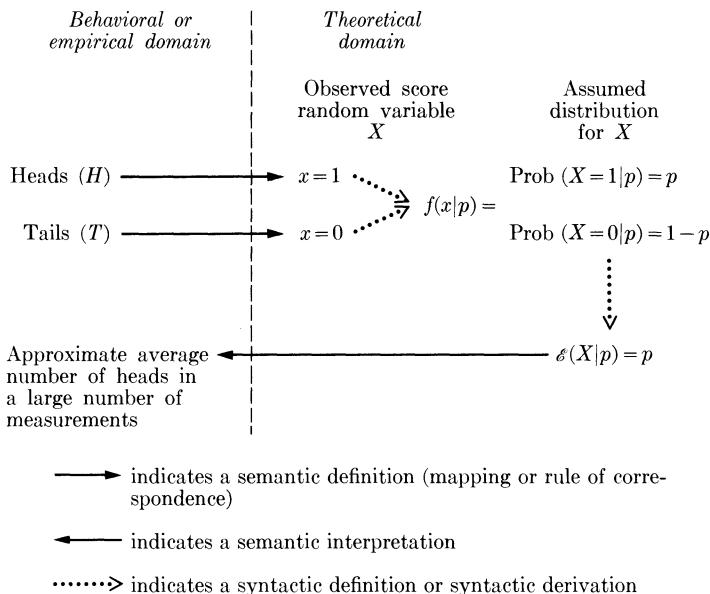


FIG. 1.3.1. Semantic and syntactic definitions.

With no loss of generality we could associate the number 1 with the symbol  $H$  and the number 0 with the symbol  $T$  and say that this formalism constitutes measurement. Others might say that this too simple rule defines something less than measurement. For our purposes, however, we shall define *measurement* to mean *a procedure for the assignment of numbers (scores, measurements) to specified properties of experimental units in such a way as to characterize and preserve specified relationships in the behavioral domain*. This assignment provides that the numbers are logically and naturally amenable to analysis by mathematical operation according to certain specified rules. In the example given above, the rule is that of inclusion and exclusion, and the property being measured is the attitude of the coin when it comes to rest. In this case the assignment of specific, distinct numbers is clearly quite arbitrary, and hence any mathematical operation on the scale of measurement which preserves the distinctness of the scores for the two possible events provides measurements consistent with the given rule. We call any such operation which preserves the distinctions made by the numerical assignment rule an *admissible* operation.

Figure 1.3.1 is a graphic presentation of the process of measurement as applied to a simple coin-tossing experiment. Tossing a specified coin results in one of two events,  $H$  or  $T$ , as described previously. The measurement procedure specifies a rule of correspondence, a mapping from this event space to a subset of the real numbers. In this case,  $H$  is mapped into the value one and  $T$  into the value zero. To characterize the relative frequency with which the events  $H$  and  $T$  will occur over repeated coin tossings, we define a random variable,

denoted by the upper-case letter  $X$ , which corresponds to the particular values denoted generically by the lower-case letter  $x$ .

The (*probability*) *distribution function*  $F(X)$  is a function which gives the probability that a random variable  $X$  will be equal to or less than any specified value  $x$ :

$$F(x) = \text{Prob} (X \leq x).$$

To every distribution function there corresponds a (*probability*) *density function*\*  $f(x)$  defined such that

$$F(x) = \int_{-\infty}^x f(t) dt$$

in the continuous case and

$$F(x) = \sum_{t \leq x} f(t)$$

in the discrete case. In the discrete case,  $f(x)$  gives the probability that  $X = x$ . The expected value and the variance of  $X$  are defined, when they exist, by

$$\mu_X \equiv \mathcal{E}X \equiv \int_{-\infty}^{\infty} xf(x) dx \quad \text{and} \quad \sigma_X^2 \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx,$$

respectively, in the continuous case, and in the discrete case by replacing the integration operation by a summation operation. These definitions are syntactic constructs having important semantic interpretations.

Generally the specification of a model involves some restrictive assumption concerning the data which the model purports to describe. For example, the Gaussian (normal) distribution has sometimes been specified as an appropriate model for mental testing applications. Such assumptions, presumably arising from the theory underlying the model, must always be made with discretion if inferences from the model are to be accurate. The simple coin-tossing experiment described above introduces relatively weak assumptions. These are (1) that the coin will never land on its edge and (2) that the result of one trial does not affect a second trial.

The definition of  $X$  as a random variable through the definition of the distribution function  $F(x)$  is a mathematical-theoretical or syntactic definition. For the experiment being considered here, the distribution function may be written as

$$F(x) = \begin{cases} 0, & -\infty < x < 0, \\ 1-p, & 0 \leq x < 1, \\ 1, & x \geq 1, \end{cases}$$

which depends on the parameter  $p$ . In fact, we might write  $F_p(x)$  or  $F(x | p)$  to emphasize the fact that each value of the parameter  $p$  gives a different dis-

---

\* In the discrete case the terms *probability function* and *probability mass function* are more common.

tribution of  $X$ . In another terminology we say that  $F$  is the conditional distribution of  $X$ , given  $p$ . Since the distribution function  $F$  is determined by  $p$ , we make inferences about  $F$  by making inferences about the value  $p$  for any given coin and possibly inferences about the distribution of  $p$  values over different coins. These inferences may then be put to use, when required, as an aid in making decisions about future coin tosses. In a language more appropriate to the analysis of responses to test items than to the analysis of coin tosses we might refer to the parameter  $p$  as the "ability" of the coin to come up heads; and indeed we might speak of a distribution of ability scores over coins. Thus by measuring the attitude of the coin over repeated tossings we can derive a measurement of its "ability to come up heads".

In the simple paradigm we have presented, each of the theoretical constructs ( $X$  and  $p$ ) has a direct semantic interpretation. This, however, is a function of the extreme simplicity of this model. In general, theoretical models involve constructs which cannot be related directly to the behavioral domain. However, in scientific theory we require that *all* constructs must be related, through syntactic definitions, to a set of basic constructs, each of which has a semantic definition. In Fig. 1.3.1 we have distinguished between semantic definitions and interpretations (indicated by  $\longrightarrow$ ) and syntactic definitions and derivations (indicated by  $\cdots\rightarrow$ ).

#### 1.4 The Place of Measurement in Psychology

The constructs of psychological theory are expressed in words. So used, these words are much more explicitly and restrictively defined, syntactically and semantically, than they are when used in their larger meanings in everyday speech or in prescientific theory. Nevertheless psychological constructs, since they are expressed in words, tend to convey broader connotative meanings.

A mathematical model differs from a purely verbal model in several respects. First, it is identified with an exact mathematical system, usually of a very high order (an algebra or calculus), by which the elementary constructs may be manipulated to facilitate deductions from the model. Also, since the mathematical model is more precise, its use avoids the confusion that results from the imprecise statement of the purely verbal model. However, because the constructs of the system are more explicitly defined, they tend to have less connotative meaning and thus, perhaps, have less conceptual richness. Finally, the mathematical model usually abstracts from and portrays only very limited aspects of the behavioral domain. The latent trait models, which are presented in the later chapters of this book, have in part been offered as a means of linking the more precise but limited mathematical models of behavior with the more broadly conceived psychological theories.

Theoretical constructs are often related to the behavioral domain through observable variables by considering the latter as *measures* or *indicants* of the former. And conversely, theoretical constructs are often abstracted from given

observable variables. We shall call an observable variable a *measure* of a theoretical construct if its expected value is presumed to increase monotonically with the construct. For example, arithmetic test scores are usually taken to be reasonable measures of arithmetic ability. Yet we know that verbal ability plays some small role in determining a person's score on most quantitative tests. (An examinee must read the directions and at times translate a word problem into quantities). A measure, however, is presumed to be primarily related to the construct being defined.

We shall call an observable variable an *indicant* of a theoretical construct if and only if its expected value is presumed to vary in a systematic way with the construct, although this relationship need not be monotonic nor need the linear correlation necessarily be nonzero. For example, an item on an attitude survey may be related in expectation to some construct by a U-shaped relationship, which could possibly yield an overall zero linear correlation between the observed variable and the construct.

### 1.5 Levels of Measurement

Certain rules for assigning numbers to events have been ordered into what are called *levels of measurement*. These levels of measurement are distinguished by the degree of complexity of the mathematical system with which they are associated, the kinds of transformations on the data which are admissible, and the kinds of operations which are meaningful in that mathematical system. We shall distinguish five levels of measurement, noting, however, that other levels are distinguishable.

The most elementary level of measurement, a classification by inclusion into mutually exclusive subsets with respect to some characteristic, as illustrated by the example in Fig. 1.3.1, has been called *nominal* (or *classificatory*) *measurement*. It yields a *nominal* (or *classificatory*) *scale*. This rule of inclusion and exclusion implies an identity or equality relationship which, with respect to some property, holds among all members of one group, but does not hold between members of different groups. The classification of persons as male or female with corresponding numerical assignment 0 or 1 constitutes a nominal measurement where the property being measured is sex. In the first group all persons have an identical value on the sex property, i.e., they are male. Clearly in this example the assigned numbers 0 and 1 are completely arbitrary and could be reversed, or, indeed, any other two sets of distinct numbers (or labels) could be chosen. Given any initial scale of measurement, any one-to-one transformation would preserve the distinction of labels and hence would in this context be termed an *admissible* (or *scale-preserving*) *transformation*.

In this very general sense, the term "measurement" has a very broad meaning. In the usage of many writers, the term "measurement" is often restricted to cases in which some stronger (i.e., more restrictive) method of scaling is employed. A nominal scale is then considered to be a classification but not a

measurement. For our purposes it is only necessary that we note that nominal measurement does not provide a "measure" in the sense defined in the previous section. In this book, however, we shall consider a strict classification or nominal scaling to be a measurement.

The second level of measurement, *ordinal measurement*, yields an *ordinal* or *rank scale*. It presupposes a set of mutually exclusive subsets as in nominal measurement, but it also presupposes that with respect to some characteristic and by some rule, an ordering relation ( $>$  or  $\geq$ ) may be established between all pairs of objects from distinct subsets. This relationship may be interpreted as "is greater than" (or "is greater than or equal to"), "is preferred to", or otherwise. If subsets are identified by real numbers, then a rule which assigns higher numbers to the more preferred subsets would be acceptable, since the ordering of the subsets would correspond to the natural ordering of the numbers. An ordinal scale gives meaning to the relative order of scale values, and any order-preserving transformation on the scale of measurement would be admissible.

The level of measurement most often specified in mental test theory is *interval measurement*, which yields an *interval scale*. This scale presupposes the ordering property of the ordinal scale but, in addition, specifies a one-to-one correspondence between the elements of the behavioral domain and the real numbers, with only the zero point and the unit of measurement being arbitrary. Such a scale assigns meaning not only to scale values and their relative order but also to relative differences of scale values. In effect, an interval scale specification establishes a distance function over all pairs of elements, with the scale unit (e.g., feet or inches) and the zero point being unspecified. Only *linear* transformations on interval scales are admissible. Such transformations preserve the equality of differences of scale scores. (See Exercise 1.1.)

A fourth level of measurement, *ratio measurement*, additionally assumes the existence of a fixed zero point and permits only multiplicative (similarity) transformations. The ratio of scale values has meaning in relation to such a scale, and hence we may speak of one unit having twice as much, or half as much, of the specified property as a second unit. Multiplicative transformations preserve the equalities of such ratios. (See Exercise 1.2.)

Finally, *absolute measurement* presupposes additionally that the unit of measurement is fixed. In this case, any transformation will destroy some property of the scale. A counting operation produces absolute measurement. Examples of ordinal, interval, ratio, and absolute measurement will be given in appropriate places later in the book.

## 1.6 The Specification of Interval Scales

Although, formally speaking, interval measurement can always be obtained by specification, such specification is theoretically meaningful only if it is implied by the theory and model relevant to the measurement procedure. At various

times in this book, however, we shall treat a measurement as having interval scale properties, although it is clear that the measurement procedure and the theory underlying it yield only a nominal or, at best, an ordinal scale. In treating data by interval methods in these cases we are, in effect, stipulating a specific distance function for our scale where the underlying measurement process and the theory supporting it have not done so. This could be considered as an arbitrary strengthening of our model. However, *from a pragmatic point of view*, the only meaningful evaluation of this procedure is one based on an evaluation of the usefulness of the resulting scale.

If we construct a test score by counting up correct responses (zero-one scoring) and treating the resulting scale scores as having interval properties, the procedure may or may not produce a good predictor of some criterion. To the extent that this scaling produces a good empirical predictor the stipulated interval scaling is justified. In Chapter 20 we shall show that in theory, under certain conditions, a better scaling procedure is available which involves the use of a weighted average of the zero-one scores. If a particular interval scale is shown empirically to provide the basis of an accurately predictive and usefully descriptive model, then it is a good scale and further theoretical developments might profitably be based on it. Thus measurement (or scaling) is a fundamental part of the process of theory construction. A major problem of mental test theory is to determine a good interval scaling to impose when the supporting psychological theory implies only ordinal properties. A major problem of theoretical psychology is to "explain" the reason for the efficacy of any particular scaling which emerges from empirical work.

Closely related to the question of scale of measurement is the question of distributional assumptions. Although much of our work is distribution free, we have included chapters on binomial and Poisson process (Poisson and negative exponential) models, and we have treated normal models in a number of chapters. Normal and Poisson models (which assume at least interval measurement) may be justified in three distinct ways: First, they may be justified by a knowledge of, or assumptions concerning, the nature of the random process that generates the random variable, as in Chapter 21. Second, they may be justified as approximations by the central-limit, Poisson-limit, and other limit theorems, as in Chapter 22. Third, they may be justified on the basis of empirical fitting, as in Chapter 23.

At times, the choice of scale for converting ordinal data to an interval scale is determined by the desire to obtain a convenient distributional form for the data. In another context, Abelson and Tukey (1959) and others have suggested that one shall choose scales so as to provide additivity of effects. In any case, distributional assumptions can only be applied to a specified scale of measurement; e.g., if  $y = \log x$  is assumed to be normally distributed, then, by definition,  $x$  must be taken to be lognormally distributed. Thus the specification of a particular distributional form for our data is equivalent to the adoption of

a particular interval or ratio scaling. It is for this reason that mathematical statistics texts pay little if any attention to problems of scale construction.

Much has been written on the foundations of measurement theory. Early work by Campbell (1920, 1957) and others on the fundamental measurement of extensive quantities laid down extremely strict requirements for the concept of measurement and severe limitations on the mathematical operations permitted to define descriptive statistics. For the most part this theoretical approach was inspired by and responsive to the methods and needs of the physical sciences. Later work by a number of writers, including Suppes and Zinnes (1963) and Luce and Tukey (1964), has greatly broadened the concept of measurement and made it much more useful to the behavioral sciences. The question of the appropriateness of any specified statistic has been recognized as one of semantic meaningfulness (e.g., Adams, Fagot, and Robinson, 1965). We believe our own very briefly stated formulation is essentially consistent with most of this later work.

### 1.7 Deterministic and Probabilistic Models

Mathematical models used in the physical sciences can, for simple problems, be put in the general form

$$x = \varphi(\boldsymbol{\theta}), \quad (1.7.1)$$

where  $x$  is considered to be a dependent (observable) variable,  $\boldsymbol{\theta}$  a vector of independent (observable or unobservable) variables, and  $\varphi$  a known function (not necessarily linear) relating  $x$  and  $\boldsymbol{\theta}$ . Given  $\boldsymbol{\theta}$ , the value of  $x$  is, in theory, exactly determined. For example, the well-known formula  $s = \frac{1}{2}g\theta^2$  purports to give the distance that a body will fall in time  $\theta$ , where  $g$  is the gravitational constant. In practice, however, the scientist generally finds that the  $\boldsymbol{\theta}$  available to him does not completely determine  $x$ ; i.e., for fixed  $\boldsymbol{\theta}$  values there is still some variation among the  $x$  values he observes. This variation is considered to be due to the effect of other independent variables not considered in (1.7.1). However, (1.7.1) may hopefully be taken as an "adequate" approximation to the true state of affairs, "adequate" in terms of the degree of accuracy that might in practice be desirable in monitoring  $x$ .

It is a characteristic of *some* branches of the physical sciences that such simple models, that depend on vector parameters  $\boldsymbol{\theta}$  which have a relatively small number of elements  $\theta_1, \theta_2, \dots, \theta_p$  (perhaps  $p \leq 3$ ), are very useful. This is true whenever for fixed  $\boldsymbol{\theta}$  and the specified function  $\varphi$  the remaining variation in  $x$  is small enough to be neglected. We say that in these cases,  $\boldsymbol{\theta}$  accounts for most of the variation in  $x$ . Equation (1.7.1) is the statement of a *deterministic model*; it asserts that  $\boldsymbol{\theta}$  determines  $x$ .

Deterministic models have found only limited use in psychology and find none in this book. The reason is simply that for problems of any real interest in mental testing we are unable to write an equation (1.7.1) such that the resid-

ual variation in  $x$  is small. Indeed, even if we use a very large number of parameters we usually are able to determine  $x$  only very inexactly. Instead, we deal with *probabilistic models* of the form

$$x = \varphi(\boldsymbol{\theta}) + e, \quad (1.7.2)$$

where  $e$ , called an error (or residual), is considered to be a composite of effects not associated with the available independent variables  $\boldsymbol{\theta}$ . Even the television meteorologist in this scientific age can give only a probabilistic statement about tomorrow's anticipated rainfall.

### 1.8 The Assumptions Underlying Test Theory Models

This book is concerned with the statement and analysis of certain mathematical models which we presume to be useful to the analysis of test score data. Each of these models is based on the axiomatic theory of probability. In addition to the fundamental axioms of probability theory, each of our models introduces certain further assumptions concerning simple properties of distributions of test scores.

If these further assumptions are valid, then the logical (mathematical) derivations from the model correctly describe more complicated relevant properties of test score data. This is just what we want. If these further assumptions do not accurately reflect the true state of affairs, then inferences from these models *may* be incorrect and possibly quite misleading. For this reason our models must be stated and employed with much discretion. We must resist the temptation to make assumptions simply because we desire to obtain certain results and formulas.

The determination of the validity or legitimacy of a mathematical model usually must be based on two separate types of study. The first type must demonstrate empirical validation of the model. This is accomplished by making predictions from a model and seeing whether these predictions are substantiated by data. The second type of study must apply logical and mathematical analysis to the assumptions of the model in such a manner that their validity or lack of validity is made more obvious and visible.

Much care must be exercised in studying the empirical validity of models. Even if a model yields quite accurate predictions in one situation, there may be little justification for inferring that it will do equally well in other situations. Also, even if a test of a model is empirically successful, we must guard against inferring the correctness of the model if the test to which we have put it is robust. For example, if we assume normal distributions for the underlying variates, we can use analysis of variance techniques with associated  $F$ -distribution tests of point hypotheses. However, it is well known that the  $F$ -test is robust against very marked deviations from these assumptions. Thus we are hardly justified in basing a test of the correctness of the normality assumptions on the relative frequency of correct  $F$ -decisions. Rather it is the case that empirical testing

of a model can have meaning only in combination with the logical analysis of the assumptions underlying the model. Mosteller (1958) has made this same point in an application of the law of comparative judgment. He graded the response percentages in an example of vegetable preferences presented by Guilford, and through suitable standardization he was able to show that five quite distinct theoretical response curves yielded nearly identical fitted values.

Often the assumptions of a model can be broken down into components whose validity, or degree of validity, is either apparent or easily investigated. We shall present one such analysis in Chapter 2. Some assumptions, and hence the models that they define, are relatively weak, in that these assumptions are obviously satisfied by most data. Other assumptions are relatively strong in that they are satisfied (even only approximately) by only very limited sets of data. It is a truism of mathematics that strong assumptions lead to strong results.

The classical test theory model of Chapters 3 through 6 is an example of a model based on relatively weak assumptions and hence one that has wide applicability. Such models may be called weak models (weak true-score models). Other models (the Poisson process models of Chapter 21 are good examples) adopt assumptions that are satisfied only under very limited conditions. Such models may be called strong models (strong true-score models). Throughout this book we shall give substantial consideration to the assumptions on which our models are based.

### **Exercises**

- 1.1. Let  $x_1, x_2, x_3$ , and  $x_4$  be values of the random variable  $X$ . Let  $Y = a + bX$  be an arbitrary linear transformation. Show that if  $x_2 - x_1 > x_4 - x_3$ , then  $y_2 - y_1 > y_4 - y_3$ .
- 1.2. Show that if  $a = 0$  in the previous exercise, and if  $x_2/x_1 = k(x_3/x_4)$ , then  $y_2/y_1 = k(y_3/y_4)$ .

### **References and Selected Readings**

- ABELSON, R., and J. W. TUKEY, Efficient conversion of non-metric information into metric information. *American Statistical Association Proceedings of the Social Statistics Section*, 1959, 226-230.
- ADAMS, E., R. F. FAGOT, and R. E. ROBINSON, A theory of appropriate statistics. *Psychometrika*, 1965, **30**, 99-128.
- BAIRD, D. C., *Experimentation, an introduction to measurement theory and experimental design*. Englewood Cliffs: Prentice-Hall, 1962.
- CAMPBELL, A., P. E. CONVERSE, W. E. MILLER, and D. E. STOKES, *Elections and the political order*. New York: Wiley, 1966.

- CAMPBELL, N. R., *Foundations of science: the philosophy of theory and experiment.* New York: Dover, 1957. (Originally appeared as *Physics: The elements*, Vol. 1. Cambridge: Cambridge University Press, 1920.)
- CARNAP, R., *Logical foundations of probability.* Chicago: University of Chicago Press, 1950. Second edition, 1962.
- GHISELLI, E. E., *Theory of psychological measurement.* Chapters 1 and 2. New York: McGraw-Hill, 1964.
- HALPERIN, M., H. O. HARTLEY, and P. G. HOEL, Recommended standards for statistical symbols and notations. *American Statistician*, 1965, **19**, 12-14.
- KEMPTHORNE, O., *An introduction to genetic statistics.* New York: Wiley, 1957.
- LUCE, R. D., and J. W. TUKEY, Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology*, 1964, **1**, 1-27.
- MOSTELLER, F., The mystery of the missing corpus. *Psychometrika*, 1958, **23**, 279-289.
- PARRATT, L. G., *Probability and experimental errors in science.* New York: Wiley, 1961.
- REIERSØL, O., Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 1950, **18**, 375-389.
- STEVENS, S. S., *Handbook of experimental psychology*, Chapter 1. New York: Wiley, 1951.
- SUPPES, P., and J. L. ZINNES, Basic measurement theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. I. New York: Wiley, 1963.
- SUTCLIFFE, J. P., A probability model for errors of classification. I. General considerations. *Psychometrika*, 1965, **30**, 73-96.
- TOPPING, J., *Errors of observation and their treatment.* London: The Institute of Physics, 1955.
- TORGERSON, W. S., *Theory and methods of scaling*, Chapters 1 and 2. New York: Wiley, 1958.
- TRUMPLER, R. J., and H. F. WEAVER, *Statistical astronomy.* Berkeley: University of California Press, 1953.

## CHAPTER 2

# THE CONSTRUCTION OF TRUE AND ERROR SCORES

### 2.1 Introduction

Although it takes a certain sophistication to recognize that the ordinary test score contains an error of measurement, few people object to this notion. An inescapable counterpart to the error of measurement does arouse objections in some quarters. If the test score  $X$  contains an error of measurement  $E$ , then clearly  $X - E$  is the measurement without the error, the true measurement, usually called the true score. The concept of the true score appears to raise some philosophical problems because often the true score cannot be directly measured. Certainly direct measurement is necessary in science; generally, however, scientists do not insist that *all* the concepts in a science must be directly measurable. Rather, as we have indicated earlier, it is sufficient that all concepts be related syntactically to other concepts that are directly measurable.

Three points of view have been taken toward true scores. One point of view (Thorndike, 1964) is that, since in the classical formulation true scores are not directly measurable, the classical notion of a true score is "mystical" and hence of no *theoretical* importance. A second point of view (Loevinger, 1957) is that, since true scores are not directly measurable and certainly can never be obtained in practice, there is little use in concerning ourselves with them or talking about them except in passing. The observed score is the only meaningful notion, and any question that cannot be answered solely by reference to observed scores is necessarily a meaningless question. From this point of view the concept of true score is of no *practical* importance. Furthermore, Loevinger (1957) contended that, as axiomatized by previous writers, the test theory model is circular. The third point of view, and our predisposition, regards the notion of true score properly defined as a conceptually useful one, which leads to many important practical results. This is not a metaphysical position; we do not advocate a theory of measurement that postulates innumerable statements that are incapable of practical verification. We use the notion of a true score because we find that it is useful theoretically and that it yields implications that can be verified in actual practice.

Let us suppose that we repeatedly administer a given test to a subject and thus obtain a measurement each day for a number of days. Further, let us assume that with respect to the particular *trait* the test is designed to measure,

the person does not change from day to day and that successive measurements are unaffected by previous measurements. Changes in the environment or the *state* of the person typically result in some day-to-day variation in the measurements which are obtained. We may view this variation as the result of errors of measurement of the underlying trait characterizing the individual, or we may view it as a representation of a real change in this trait. What we consider to be a change in trait and what we consider to be an irrelevant fluctuation, and hence what we do or do not consider to be error, depends primarily on how we define true score.

It is worthwhile, initially, to distinguish three different conceptions of true score and to make clear our reasons for selecting one of these as the basis for our work. The first of these is most closely related to the common usage of the term. For a specified measurement  $g$  and person  $a$ , we might hypothesize that the person has some unique naturally defined "true" value on the measurement being taken. We denote this true value by  $\tau_{ga}$ , and refer to it as the individual's *true score*. Any particular measurement yields a value  $x_{ga}$  which in general differs from  $\tau_{ga}$  for many different reasons. The value  $x_{ga}$  is observable, the constant  $\tau_{ga}$  is not. This simple Platonic conception (Sutcliffe, 1965) of true score seems to underlie much of the thinking and writing in the classical theory of errors, and it is the conception criticized by Thorndike (1964). In the physical sciences, this conception is a very reasonable one. If conditions are precisely specified, it seems correct to speak of the true velocity of light, the true weight of a bag of potatoes, and the true distance between two points, assuming of course that a measurement procedure and hence a scale of measurement have been specified. However, this conception of true score does not provide a satisfactory axiomatic basis for psychological theories, since these theories are typically based on unexplicated, inexact constructs.

A second conception of true score, the one most widely offered in the test theory literature, parallels the relative frequency axiomatization of probability theory due to von Mises (see Carnap, 1962). If successive measurements are taken in the manner described above and if the average value of these observations is determined, then (under very mild distributional assumptions) that average value will converge, with probability one, to a constant. This constant may be denoted by  $\tau_{ga}$  and called the true score of subject  $a$  on measurement  $g$ . Paralleling this mathematical result, the true score has at times been defined as the limiting value of the (average) score that a person would obtain as the length of the test increased; that is, as the average score on a test consisting of infinitely many replications. The von Mises axiomatization of probability theory has proved unsatisfactory for technical reasons (see Lindley, 1965, p. 5) and it is little used in contemporary work. A similar consideration makes it undesirable to define true score as a limiting average.

The classical test theory model which we shall use throughout the first part of this book is based on a particular, mathematically convenient and conceptually useful, definition of true score and on certain basic assumptions

concerning the relationships among true and error scores. Specifically it is assumed that corresponding true and error scores are uncorrelated and that error scores on different measurements are also uncorrelated. In discussing these assumptions Guttman (1945), Gulliksen (1950), and others have indicated that these assumptions can be thought of more as definitions than as axioms. In this chapter, following Novick (1966), we give a detailed demonstration of how true and error scores, under very general conditions, may be *constructed* (defined) so that they satisfy the assumptions of the classical theory. These definitions are purely syntactic in nature, though we provide parallel semantic interpretations of these concepts by relating them to both the Platonic and limiting-value conceptions of true and error score. Our formulation generalizes the classical model. It contains it as a special case and also serves as a basis for much more general models developed in Chapters 8 through 11.

## 2.2 The Distribution of Measurements on a Fixed Person

Let us think in terms of some well-defined population  $\mathcal{P}$  of persons (experimental units) and some well-defined set  $\mathcal{I}$  of measurements corresponding to a set of tests or test items. In theory,  $\mathcal{P}$  and  $\mathcal{I}$  may contain a (countably) infinite number of elements or a finite number of elements, possibly only a single one. We shall use the subscript  $a$  to refer to the  $a$ th person and the subscript  $g$  to refer to the  $g$ th measurement. The expressions  $a \in \mathcal{P}$  and  $g \in \mathcal{I}$  mean that  $a$  is a member of the set  $\mathcal{P}$  and  $g$  is a member of the set  $\mathcal{I}$ . Taking a measurement  $g$  on person  $a$  results in a numerical value which we denote by  $x_{ga}$ . This value depends on the particular measurement situation, and it is only one of a number of possible values which might be obtained. Therefore we conceive the score  $x_{ga}$  to be the realization of a random process, and we associate the value  $x_{ga}$  with a real-valued random variable  $X_{ga}$  defined on the set of all possible values  $x_{ga}$  that might be observed. Under repeated measurement of a single examinee with a given test, various sets of  $x_{ga}$  values will occur more or less frequently, the frequency depending on the effect of the various conditions on the values actually observed and on the relative frequency with which these conditions occur under some well-defined random sampling procedure.

To provide a first semantic interpretation for this process, let us begin with the explanation given by Lazarsfeld (1959).

Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favor of the United Nations; suppose further that after each question we "wash his brains" and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfavorable answer. Having gone through this procedure many times, we then compute the proportion of times Mr. Brown was in favor of the United Nations. . . .

There is one interesting consequence of this version of the probability notion. A specific Mr. Brown, for instance, might feel that it is his duty to be in favor of the United Nations. Therefore, if he is asked a question when he is sober, his probability—or, if you please, his propensity—to be in favor of the United Nations might be rather

high. Under the influence of alcohol, however, his hostility to the international organization might come out. Therefore, his probability under the influence of alcohol could be different than his probability if he were sober. This is an idea which is often used in the parlance of daily life. A man can drive "60 miles an hour" and at the next moment if a policeman is around, slow down to "40 miles an hour". What we call miles per hour is not what a man actually drives within an hour, but what he would drive if certain specified conditions were to prevail. Thus, we have a concept of probability which can apply to a single individual: furthermore, this probability or propensity itself can be different under various conditions. (pp. 493-494)\*

Most students taking college entrance examinations are convinced that how they do on a particular day depends to some extent on "how they feel that day". A student who receives scores which he considers surprisingly low often attributes this unfortunate circumstance to a physical or psychological indisposition or to some more serious *temporary* state of affairs not related to the fact that he is taking the test that day. To provide a mathematical model for such cyclic variations, we conceive initially of a sequence of independent observations, as described by Lazarsfeld, and consider some effects, such as the subject's ability, to be constant, and others, such as the transient state of the person, to be random. We then consider the distribution that might be obtained over a sequence of such statistically independent measurements if each were governed by the *propensity distribution*  $F_{ga}(x_{ga})$  of the random variable  $X_{ga}$ . The function  $F_{ga}(x_{ga})$  is a cumulative distribution function defined over repeated statistically independent measurements on the same person. Notationally, for fixed  $g$  and  $a$ ,  $F_{ga}(x_{ga}) = \text{Prob}(X_{ga} \leq x_{ga})$  over repeated measurements  $X_{ga}$ . The propensity distribution is a hypothetical one because, as we noted in Chapter 1, it is not usually possible in psychology to obtain more than a few independent observations. Even though this distribution cannot in general be determined, the concept will prove useful.

### 2.3 True Score as an Expectation

A mathematically convenient way of defining a true score is as the expected observed score with respect to the propensity distribution of a given person on a given measurement. Therefore we state

**Definition 2.3.1.** *The true score  $\tau_{ga}$  of a person  $a$  on measurement  $g$  is defined as the expected value of the observed score; that is,*

$$\tau_{ga} \equiv \mathcal{E} X_{ga}, \quad (2.3.1)$$

where  $X_{ga}$  is a random variable taking values  $x_{ga}$ ,  $\mathcal{E}$  denotes expectation with respect to the propensity distribution, and  $\tau_{ga}$  is a constant, the expected value of  $X_{ga}$ .

---

\* S. Koch (Ed.), *Psychology: a study of a science*, Copyright 1959. McGraw-Hill Book Company. Used by permission.

We assume here that this expectation exists, that is, that it has a finite value. In test theory this assumption of finite expectation is relatively weak from both a theoretical and a practical point of view. For suppose that a random variable has infinite expectation; it is then always possible to transform this random variable, using only a strictly increasing transformation, to another which is defined on the interval from zero to one. Since any random variable defined on a finite interval has finite moments of every (positive) order, the transformed random variable will certainly have finite expectation. In practice, mental tests generally have bounded scores and hence have finite moments of every order.

In effect the transformation suggested above involves the determination of a scale of interval measurement from a scale of ordinal measurement. As we have indicated in Section 1.5, such ordinal scales are defined up to a monotone transformation, and therefore such change of scale of measurement is admissible. Indeed the choice of observed-score scales might well be made in part to depend on the existence of the moments of the associated distribution.

We note two important things about true scores. First, it is clear that the true-score scale is determined by the observed-score scale through Eq. (2.3.1). Second, true scores are not directly measurable. They are, however, related theoretically to observed scores (constructs which are directly measurable) through the relation (2.3.1), which is a *syntactic* (or mathematical) *definition* of true score. In Section 5.4 we shall use the law of large numbers to show that under certain assumptions about the real world, this true score may be *interpreted* semantically as the average score that the person would obtain on infinitely many independent repeated measurements (an unobservable quantity). The validity of the law of large numbers, however, depends on the existence of the expected value. Thus the expected value is mathematically a more basic concept than the limiting value and hence a better choice for our syntactic definition of true score.

## 2.4 The Construction of the Error Random Variable for a Fixed Person

The discrepancy between observed value and true value has usually been referred to as the error of measurement. For our purposes, we formally define the random variable  $E_{ga}$  by the equation

$$E_{ga} \equiv X_{ga} - \tau_{ga}, \quad (2.4.1)$$

where  $E_{ga}$  and  $X_{ga}$  are random variables and where  $\tau_{ga}$  is a constant, the expected value of  $X_{ga}$ . Formally we have

**Definition 2.4.1.** *The error-score random variable  $E_{ga}$  is the difference  $X_{ga} - \tau_{ga}$  between the observed-score random variable and the constant  $\tau_{ga}$ , the true score.*

Because of the *constructed* linear relation (2.4.1), the (linear) correlation between  $X_{ga}$  and  $E_{ga}$  with respect to the propensity distribution is unity. Also, since  $\mathcal{E}X_{ga} = \tau_{ga}$ , we have

$$\mathcal{E}E_{ga} = \mathcal{E}(X_{ga} - \tau_{ga}) = \tau_{ga} - \tau_{ga} = 0. \quad (2.4.2)$$

The true and error scores defined above are not those primarily considered in test theory (see Section 8.2). They are, however, those that would be of interest to a theory that deals with individuals rather than with groups (counseling rather than selection) or with quota-free selection (Cronbach and Gleser, 1965). In the following sections we shall construct the random variables with which test theory is most concerned.

We note that this construction of  $\tau_{ga}$  and  $E_{ga}$  may be carried out separately and independently for each person  $a$ . We need not assume that the propensity distribution variances

$$\sigma^2(E_{ga}) = \sigma^2(X_{ga}) = \mathcal{E}(X_{ga} - \tau_{ga})^2 \quad (2.4.3)$$

are equal for different persons. Thus we allow the possibilities that some persons' responses are inherently more consistent than those of others, and that we are able to measure some persons' responses more accurately than others'. An actual construction of true and error-score random variables is outlined in Exercise 2.17.

## 2.5 The Random Selection of Persons

We have to this point dealt with hypothetical replications (described by the propensity distribution) on one person. Primarily, test theory treats individual differences or, equivalently, the distribution of measurements over people. Therefore let us now consider a somewhat different measurement procedure. Suppose that we *randomly* select a person  $a$  from the population  $\mathcal{P}$  and administer the test  $g$  under standardized conditions. In theory this random selection and administration can be done repeatedly, and we can thus generate an observed-score random variable  $X_{g*}$  taking specific values  $x_{g*}$ . Replacement of the subscript  $a$  by the subscript  $*$  will denote random selection of a person for each measurement rather than the specific selection of a specific person. In most contexts we shall denote these observed values by  $x_{ga}$  since, after sampling, the person  $a$  is determined.

We may also define the (cumulative) distribution function  $F_{g*}(x)$  which, in the population  $\mathcal{P}$ , gives the probability that  $X_{g*} \leq x_{g*}$ . Semantically  $F_{g*}$  gives the proportion of measurements, obtained as described above, which are less than  $x_{g*}$ . Thus  $F_{g*}(x)$  pertains to different randomly selected persons. We denote the expectation and variance of  $X_{g*}$  by  $\mathcal{E}(X_{g*})$  and  $\sigma^2(X_{g*})$ , respectively.

In a similar manner we may define the random variables  $T_{g*}$  and  $E_{g*}$ , which take values  $\tau_{ga}$  and  $e_{ga}$ , over random sampling of persons. We shall denote the (cumulative) distribution functions of these random variables by  $G_{g*}(\tau)$  and  $H_{g*}(e)$ . Whether persons are selected randomly or by specification, the experimental (sampling) unit is the person  $a$ , and a sample of size one consists of the triplet  $(x_{ga}, \tau_{ga}, e_{ga})$ , of which only the first element is observable.

Suppose  $\mathcal{P}$  contains three persons, having true scores  $\tau_{g1} = -1$ ,  $\tau_{g2} = 0$ , and  $\tau_{g3} = 1$ . Suppose  $E_{ga} = \pm a$ ,  $a = 1, 2, 3$ , each with probability  $\frac{1}{4}$ , and  $E_{ga} = 0$  with probability  $\frac{1}{2}$ . Then  $T_{g*}$  is a random variable taking values  $\tau_{ga}$  of

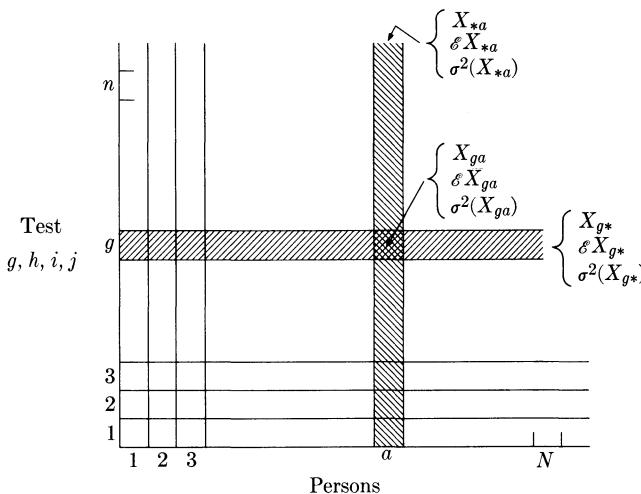


FIG. 2.5.1. Schematization of sampling procedures.

—1, 0, and 1, each with probability  $\frac{1}{3}$ , and  $E_{g*}$  is a random variable taking values  $-3, -2, -1, 1, 2$ , and 3 with probability  $\frac{1}{12}$  and taking the value 0 with probability  $\frac{1}{2}$ . The random variable  $X_{g*}$  then takes the following values  $x_{ga}$  with the indicated probabilities:

$$\begin{array}{c|cc} x_{ga} & \text{Prob } (X_{g*} = x_{ga}) \\ \hline -2 & \frac{3}{12} \\ -1 & \frac{2}{12} \end{array} \quad \begin{array}{c|cc} x_{ga} & \text{Prob } (X_{g*} = x_{ga}) \\ \hline 0 & \frac{3}{12} \\ 1 & \frac{2}{12} \end{array} \quad \begin{array}{c|cc} x_{ga} & \text{Prob } (X_{g*} = x_{ga}) \\ \hline 2 & \frac{1}{12} \\ 4 & \frac{1}{12} \end{array}$$

In all cases we shall be concerned with sampling with replacement and we shall usually be concerned either with subpopulations that contain infinitely many people or that contain only a single person. In the former case the probability of a person being sampled twice is zero, and hence we may speak of sampling over persons. In the latter case the same person is sampled each time a measurement is taken and the general sampling model reduces to sampling with respect to the propensity distribution for a fixed person. For this reason it is not necessary to make the expectation over the propensity distribution explicit. However, in Section 2.12 we shall show that sampling with respect to the propensity distribution is realized operationally by specific kinds of repeated measurements called *replications*. At that point we shall introduce an indexing subscript  $k$  to denote the  $k$ th replication and thereafter denote the expectation  $\mathcal{E}E_{ga}$  for fixed  $g$  and  $a$  as the expectation  $\mathcal{E}_k E_{gak}$  over replications. In Chapters 7 through 10 this explicit denotation of the expectation over replications will prove convenient to distinguish this type from expectations taken over persons or over tests.

The sampling rules for the random variables  $X_{ga}$  and  $X_{g*}$  are schematized in Fig. 2.5.1. The random variable  $X_{ga}$  is defined over repeated sampling, with

replacement, from the cell  $(g, a)$ . The random variable  $X_{g*}$  is defined over repeated sampling, with replacement, from row  $g$ . We extend these ideas by defining random variables  $X_{*a}$  and  $X_{**}$ . These are defined over repeated measurements, with replacement, from column  $a$  for  $X_{*a}$  and over the entire  $(g, a)$  grid for  $X_{**}$ . These last two random variables are the subject of an intensive study in Chapters 8 and 11.

## 2.6 Construction of the Linear Model

For the random variables  $X_{g*}$ ,  $T_{g*}$ , and  $E_{g*}$  constructed in Section 2.5 by random sampling of people, we have

### Theorem 2.6.1

$$X_{g*} = T_{g*} + E_{g*} \quad (2.6.1)$$

in  $\mathcal{P}$  or any subpopulation of  $\mathcal{P}$ .

*Proof.* Since Eq. (2.4.1) holds for an arbitrarily selected person  $a$ , it is apparent that the equation  $x_{ga} = \tau_{ga} + e_{ga}$  holds for a randomly selected person from  $\mathcal{P}$  or any subpopulation of  $\mathcal{P}$ . Hence it is clear that the error random variable  $E_{g*}$  is just the difference between the random variables  $X_{g*}$  and  $T_{g*}$ .  $\square$

If the subpopulation of  $\mathcal{P}$  contains only the single person  $a$ , then Eq. (2.6.1) reduces to Eq. (2.4.1). In obtaining the linear relation (2.6.1), no assumption was required other than that of the existence of  $\tau_{ga} \equiv \mathcal{E}X_{ga}$ . Hence the linear relation (2.6.1) involves no assumption of linearity whatsoever, but rather follows directly from the definitions of the random variables  $X_{g*}$ ,  $T_{g*}$ , and  $E_{g*}$ .

The sampling rule underlying Eq. (2.6.1), the various expectation operations associated with these random variables, and the notation we shall use to denote these operations require careful statement. Equation (2.6.1) states a linear relation involving random variables described by the following model. We have a defined population  $\mathcal{P}$  of persons. We randomly select one person from this population and observe his performance on a test (item). This person has a fixed true score. On this occasion his particular error score and hence his particular observed score are determined by the propensity distribution. Thus each observation resulting from the random selection of a person generates three scores: the observed score, the true score, and the error score. Of course, only the observed score is observable by the experimenter. These three scores are characterized jointly by the random variables  $X_{g*}$ ,  $T_{g*}$ , and  $E_{g*}$ , as defined above, and the measurement is sometimes denoted by  $(X_{g*}, T_{g*}, E_{g*})$  or more simply by  $(X, T, E)$ .

In Eq. (2.4.2) we introduced the symbol  $\mathcal{E}E_{ga}$  to indicate the expected value with respect to the propensity distribution (i.e., over repeated measurements) of the random variable  $E_{ga}$  for fixed  $g$  and fixed  $a$ . We now introduce the notation  $\mathcal{E}_a E_{ga}$ , or alternatively and equivalently  $\mathcal{E}E_{g*}$ . Whenever we use either of these two symbols we shall mean that a person  $a$  is selected *randomly* and the ex-

pected value with respect to the propensity distribution of the resulting random variable is then obtained. Another useful notation for this dual expectation is  $\mathcal{E}_a \mathcal{E} E_{ga}$ . This use of the expectation operator  $\mathcal{E}_a$  to denote a random sampling operation is standard in sampling theory. We note for future reference that

$$\mathcal{E}_a E_{ga} = \mathcal{E} E_{g*} = \mathcal{E}_a E_{gak} = \mathcal{E} E_{ga} = \mathcal{E}_a \mathcal{E}_k E_{ga}. \quad (2.6.1a)$$

As further examples of this type of notation, consider the expressions  $\sigma^2(E_{g*})$  and  $\mathcal{E}_a \sigma^2(E_{ga})$  for the variance of the error score for a randomly selected person and the average over people of the variance of the error scores for a fixed person, respectively. In Exercise (2.7) the reader is asked to prove that  $\sigma^2(E_{g*}) < \infty$  and  $\sigma^2(E_{ga}) < \infty$  for all  $a$ . In Theorem 2.7.1 we shall show that  $\mathcal{E} E_{g*} = 0$ . Hence we may then conclude that

$$\sigma^2(E_{g*}) = \mathcal{E} E_{g*}^2 \quad \text{and} \quad \mathcal{E}_a \sigma^2(E_{ga}) = \mathcal{E}_a \mathcal{E} E_{ga}^2.$$

But the definition of the operator  $\mathcal{E}_a$  implies that

$$\mathcal{E}_a \mathcal{E} E_{ga}^2 = \mathcal{E} \mathcal{E}_a E_{ga}^2 = \mathcal{E} E_{g*}^2,$$

and hence

$$\sigma^2(E_{g*}) = \mathcal{E}_a \sigma^2(E_{ga}). \quad (2.6.2)$$

*Thus the error variance over persons is equal to the average, over persons, of the error variance within persons.*

It is important to note that, for example,  $\sigma^2(E_{ga})$  for fixed  $g$  and  $a$  is a constant. But  $\sigma^2(E_{ga})$  may vary over a population of people or tests or both. Hence, over people, or tests, or tests and people,  $\sigma^2(E_{ga})$  is the value taken by an appropriately defined random variable. Thus we may speak of  $\mathcal{E}_a \sigma^2(E_{ga})$ , by which we shall mean the expected value of  $\sigma^2(E_{ga})$  over people.

When relating  $\sigma^2(E_{g*})$  to  $\sigma^2(E_{ga})$  and when making a great many other similar comparisons, one basic and extremely useful formula finds repeated use. For arbitrary random variables  $Z$  and  $Y$  we have

**Theorem 2.6.2**

$$\sigma^2(Z) = \mathcal{E} \sigma^2(Z | Y) + \sigma^2[\mathcal{E}(Z | Y)]. \quad (2.6.3)$$

The proof of this theorem is given by Freeman (1963, pp. 54–57) and in many other texts. In effect this formula is a generalization of the usual analysis of variance breakdown of total variance into the sum of (1) average within-class variance and (2) among-class variance. As an example, for the random variables  $E_{g*}$  and  $E_{ga}$  we have

$$\begin{aligned} \sigma^2(E_{g*}) &= \mathcal{E}_a \sigma^2(E_{ga}) + \sigma_a^2(\mathcal{E} E_{ga}) \\ &= \mathcal{E}_a \sigma^2(E_{ga}), \end{aligned} \quad (2.6.4)$$

since  $\mathcal{E} E_{ga} = 0$  for all  $a$ .

## 2.7 Derivation of the Usual Assumptions of the Classical Model

As we proceed with the linear model constructed in Section 2.6 we require first the additional simple assumption that the variance of observed scores,  $\sigma^2(X_{g*})$ , be finite.\* Solely to avoid triviality,  $\sigma^2(X_{g*})$ ,  $\sigma^2(T_{g*})$ , and  $\sigma^2(E_{g*})$  are in this book usually assumed to be strictly greater than zero although at times we shall be interested for expository purposes in the extreme cases where one or the other is zero. The question of the finiteness of  $\sigma^2(T_{g*})$  we leave as an exercise, 2.7, for the reader. Further, if the  $(r+1)$ -moment of a distribution is finite, then the  $r$ th moment must also be finite,  $r = 1, 2, \dots$ ; hence the expected values of the observed scores and true scores,  $\mathcal{E}(X_{g*})$  and  $\mathcal{E}(T_{g*})$ , must be finite since  $\sigma^2(X_{g*})$  is finite (Kendall and Stuart, 1958, p. 63). From these simple assumptions we shall now show that the expected error score is zero, that true scores are uncorrelated with error scores, and that  $\mathcal{E}(X_{g*}) = \mathcal{E}(T_{g*})$ .

In Section 2.2 we quoted Lazarsfeld's assumption that a brainwashing process occurs between repeated observations. Lazarsfeld was supposing that the measuring procedure does not affect the ability or true score of the person measured or the propensity distribution in future measurements. Here we adopt the same sort of notion by assuming that, for "distinct" measurements  $g$  and  $h$ ,  $X_{ga}$  and  $X_{ha}$  (or equivalently  $E_{ga}$  and  $E_{ha}$ ) are independently distributed for each  $a$  in  $\mathcal{P}$ .

In Sections 2.10 and 2.11 we shall discuss in greater detail the nature of the assumptions underlying the following basic theorem and the extent to which the given assumption can be further relaxed.

**Theorem 2.7.1.** For the true-score and error-score random variables  $T_{g*}$  and  $E_{g*}$  constructed in Sections 2.3 through 2.6,

- i) *the expected error score is zero,*

$$\mathcal{E}E_{g*} = 0; \quad (2.7.1a)$$

- ii) *the correlation between true and error scores is zero,*

$$\rho(E_{g*}, T_{g*}) = 0; \quad (2.7.1b)$$

- iii) *the correlation between the error score on one measurement and the true score on a second is zero,*

$$\rho(E_{g*}, T_{h*}) = 0, \quad (2.7.1c)$$

where  $\rho$  denotes a correlation. Under the further assumption that  $X_{ga}$  and  $X_{ha}$  are independently distributed for each  $a$ ,

- iv) *the correlation between errors on distinct measurements is zero,*

$$\rho(E_{g*}, E_{h*}) = 0. \quad (2.7.1d)$$

---

\* This assumption *does not follow* from the assumption that  $\sigma^2(X_{ga}) < \infty$  for all  $a$ . However, it may be justified by the discussion given in Section 2.3.

*Proof.* The proof of this theorem is simple and instructive.

$$\text{i) } \mathcal{E}E_{g*} = \mathcal{E}_a(\mathcal{E}E_{ga}) = \mathcal{E}_a 0 = 0,$$

where the expectation in parentheses is taken over the propensity distribution, or equivalently over replications (see Section 2.12). We have here made explicit use of the double source of variation in  $E_{g*}$  so that its expectation may be re-written as a double expectation, first taken with respect to the propensity distribution and then taken over people.

ii) Now let  $\beta \equiv \beta(E_{g*} | \tau_{ga})$  be the linear regression coefficient of  $E_{g*}$  on  $\tau_{ga}$  and let  $\rho_{ET}$  and  $\sigma_E$  and  $\sigma_T$  be the corresponding correlation and standard deviations. Then  $\beta$  is equal to  $\sigma_E \rho_{ET} / \sigma_T$  and is the slope of the regression line of  $E_{g*}$  on  $\tau_{ga}$ . Since  $\mathcal{E}E_{ga} = 0$  for each  $(g, a)$ , it will also be zero for a randomly selected  $a$  from *any* subset of  $\mathcal{P}$ . In particular, it will be zero for every  $a$  such that  $\tau_{ga}$  is any specified constant. Thus we have shown that the regression function  $\mathcal{E}(E_{g*} | \tau_{ga})$  has the constant value 0 for each  $\tau_{ga}$ , that is, the expected value of  $E_{g*}$  in the subpopulation of people with true scores  $\tau_{ga}$  is zero. The regression of  $E_{g*}$  on  $\tau_{ga}$  is thus linear, so that the regression function and the linear mean squared error regression function are identical and each has slope  $\beta = 0$ . Since  $\rho(E_{g*}, T_{g*}) = \beta \sigma_T / \sigma_E = 0$ , the desired result, which is indeed much weaker than the obtained result, follows.

- iii) The proof of this part of the theorem is left as an exercise for the reader.
- iv) Similarly for *fixed* measurements  $g$  and  $h$ , under the assumption that  $X_{ga}$  and  $X_{ha}$  (and hence  $E_{ga}$  and  $E_{ha}$ ) are independently distributed (for fixed  $a$ ), we have  $\mathcal{E}(E_{g*} | e_{ha}) = 0$  for all  $e_{ha}$ , and hence  $\rho(E_{g*}, E_{h*}) = 0$ .  $\square$

Now applying the methods described in the previous section we have

**Theorem 2.7.2**

$$\mathcal{E}X_{g*} = \mathcal{E}(T_{g*} + E_{g*}) = \mathcal{E}T_{g*}; \quad (2.7.2)$$

*the mean observed score is equal to the mean true score.*

From the work in Section 2.6, we have

**Theorem 2.7.3**

$$\text{i) } \sigma^2(E_{g*}) = \mathcal{E}E_{g*}^2 = \mathcal{E}_a E_{ga}^2 = \mathcal{E}_a \mathcal{E}E_{ga}^2 = \mathcal{E}_a \sigma^2(E_{ga}), \quad (2.7.3a)$$

and since

$$E_{ga} = X_{ga} - \tau_{ga}$$

where  $\tau_{ga}$  is a constant, we have

$$\sigma^2(E_{ga}) = \sigma^2(X_{ga});$$

hence

$$\text{ii) } \sigma^2(E_{g*}) = \mathcal{E}_a \sigma^2(X_{ga}). \quad (2.7.3b)$$

However, note that in general

$$\sigma^2(X_{g*}) \neq \mathcal{E}_a \sigma^2(X_{ga}), \quad (2.7.3c)$$

for

$$\begin{aligned} \sigma^2(X_{g*}) &= \mathcal{E}_a \sigma^2(X_{ga}) + \sigma_a^2[\mathcal{E}(X_{ga})] \\ &= \mathcal{E}_a \sigma^2(X_{ga}) + \sigma^2(T_{g*}). \end{aligned}$$

The last term is in general assumed to be greater than zero.

Furthermore, from the proof of (ii) of Theorem 2.7.1, we may separately state the following result:

**Theorem 2.7.4.** For all  $\tau_{ga}$ ,

$$\mathcal{E}(E_{g*} | T_{g*} = \tau_{ga}) = 0; \quad (2.7.4)$$

*the expected error score, for sampling over people in the subpopulation of people having any fixed true score, is zero.*

The results of Theorem 2.7.1 which concern the random variables  $X_{g*}$ ,  $T_{g*}$ , and  $E_{g*}$  define the *classical test theory model*. We have shown that these statements are not assumptions but properties derived from much simpler assumptions. From another point of view it may be said that these assumptions are in fact a definition of what we mean by true and error scores. In Chapters 3 through 6 we shall study the implications of this simple model. However, the approach outlined in the present chapter also forms the basis for more general weak true-score models (see Chapters 8 through 11) and indeed is basic to all of true-score theory.

## 2.8 What is Error?

Though we have indicated that we shall employ only probabilistic models in this book, our point of view is, in a sense, consistent with a completely deterministic view of nature. Under the model developed in the preceding sections, we consider the error-score random variable  $E_{ga}$  to be a disturbance that is due to a composite of a multitude of factors not controlled in the measurement procedure. Error variation can sometimes be reduced by controlling additional factors in the environment. It is partly for this reason that the conditions under which tests are given must be controlled. From this point of view, for a specified person and a population of experimental conditions, the error random variable is a perfectly proper (nondegenerate) random variable, and the true score is a constant (a degenerate random variable). *Thus the error random variable and the true score are determined by the experiment, not by some hypothetical state of affairs.*

Just as we distinguished between the Platonic and operationally related conceptions of true score, we must distinguish between such conceptions of error score. Our method of constructing true and error scores has provided an

analytic decomposition of  $X_{g*}$  into the sum of two uncorrelated components, a true score and an error score, each of which is syntactically defined in terms of  $X_{g*}$ . From this variance analysis point of view, the error random variable might more appropriately be called the *residual random variable* and the variance of this random variable the *residual variance*.

The error random variable may have a number of components associated directly with the physical measurement process, others associated with the environment in which the measurements are taken, and still others associated with temporal changes in the person. Differing degrees of control and randomization of these components define different true scores and hence different residual (error) scores. For each definition of true score, of course, we have a different error score. Just what is included in the error score depends entirely on the conditions under which measurements are made. In the physical sciences relatively simple controls on a small number of conditions often eliminate, for all practical purposes, the residual variation (error variance) in an experiment.

The model we have constructed is a special case of the model used in variance components analyses (model II: the random components model). The true scores of the classical test theory model play the part of main effects and the error scores, the part of within-group variation. In Chapters 7 through 9 we shall adapt the methods of variance components analysis to the problem of estimating the means and variances of the true- and error-score distributions.

## 2.9 The Many Concepts of True Score\*

While the true score as an expectation is the concept of true score used throughout most of this book, it is by no means the only useful definition of true score. In this section we shall briefly discuss other possible approaches.

In Section 2.1 we referred to a Platonic conception of true score and simply noted that this concept usually does not coincide with the conception of true score adopted in this book. The terminology *Platonic true score* was introduced into the psychometric literature by Sutcliffe (1965) as a contrast to his operationally related basis for a theory of classification.† The Platonic character of the concept of true score as presented in the classical test theory literature had previously been criticized by R. L. Thorndike (1964) and others. To distinguish between Platonic and operationally related definitions of true score, we shall adapt an example and a development given by Sutcliffe.

We are concerned with the work of a chicken-sexer and the accuracy of his classification of pullets and cockerels. We shall not initially concern ourselves with repeated measurements (classifications), but rather limit our attention to the sampling of chickens. Now each of the chickens is truly either a pullet or a

---

\* Reading of this section may be omitted without loss of continuity.

† We may describe Sutcliffe's approach as one of defining the true score to be the *mode* of the propensity distribution (whereas we have defined it to be the *expected value*), although Sutcliffe restricts himself to a discrete, finite sample space.

Table 2.9.1

Distribution of true and observed  
pullet-cockerel classifications

		Observed score ( $x$ )		$p_{1\cdot}$		
		(Pullet)	(Cockerel)			
True score ( $\tau$ )	0	1		$p_{0\cdot}$		
	(Cockerel)	1	$p_{10}$	$p_{11}$		
		0	$p_{00}$	$p_{01}$		
			$p_{\cdot 0}$	$p_{\cdot 1}$		

cockerel. After a brief examination the chicken-sexer classifies each chicken as either a pullet or a cockerel. His assigned classifications (observed scores) do not always coincide with the true classifications (true scores); he classifies some pullets as cockerels and some cockerels as pullets.

The level of measurement here is clearly nominal both with respect to true scores and with respect to observed scores. Therefore we may score (or code) the pullet classification as 0 and the cockerel classification as 1. The observed-score random variable  $X$  and true-score random variable  $T$  thus take values 0 and 1. Let  $p_{ij}$ ,  $i = 0, 1$  and  $j = 0, 1$ , be the probability of a randomly selected chicken being truly  $i$  but classified as  $j$ . Thus we may say that  $p_{01}$  is the proportion of chickens (in the chicken population) which are truly pullets but which are classified as cockerels by the chicken-sexer. Let  $p_{0\cdot}$ ,  $p_{1\cdot}$ ,  $p_{\cdot 0}$ , and  $p_{\cdot 1}$  be the proportions of true pullets, true cockerels, observed pullets, and observed cockerels in the population. We summarize these definitions in Table 2.9.1. The marginal probabilities are defined by

$$p_{1\cdot} = p_{10} + p_{11}, \quad p_{0\cdot} = p_{00} + p_{01}, \quad p_{\cdot 0} = p_{00} + p_{10}, \quad p_{\cdot 1} = p_{01} + p_{11}.$$

The conditional probability of a true pullet being classified as a cockerel is  $p_{01}/p_{0\cdot}$  and the conditional probability of a true cockerel being classified as a pullet is  $p_{10}/p_{1\cdot}$ .

We may also consider Table 2.9.2, which compares true scores ( $\tau$ ) and error scores ( $e$ ), where the error scores are defined by the usual equation  $E = X - T$ .

Assuming that the necessarily nonnegative quantities  $\sigma^2(E)$  and  $\sigma^2(T)$  are strictly positive (i.e., nonzero), we can compute the correlation

$$\begin{aligned} \rho(T, E) &= \frac{\sigma(T, E)}{\sigma(E)\sigma(T)} = \frac{[\mathcal{E}(ET) - \mathcal{E}(E)\mathcal{E}(T)]}{\sigma(E)\sigma(T)} \\ &= \frac{-p_{10} + p_{1\cdot}(p_{10} - p_{01})}{\sqrt{p_{10} + p_{01} - (-p_{10} + p_{01})^2} \sqrt{p_{1\cdot} - p_{\cdot 1}^2}}. \end{aligned}$$

**Table 2.9.2**

Distribution of true and error  
pullet-cockerel classifications

		Error score ( $e$ )			
		-1	0	+1	
True score ( $\tau$ )	(Cockerel) 1	$p_{10}$	$p_{11}$	0	$p_{1\bullet}$
	(Pullet) 0	0	$p_{00}$	$p_{01}$	$p_{0\bullet}$
		$p_{10}$	$p_{11} + p_{00}$	$p_{01}$	

If we rewrite the numerator as  $-p_{10}(1 - p_{1\bullet}) - p_{01}p_{1\bullet}$ , it is clear that

$$-1 \leq \rho(T, E) < 0.$$

The quantity  $\rho(T, E)$  is undefined whenever all classifications are correct; however, the limiting value in this case is 0.

We observe that the result  $-1 \leq \rho(T, E) < 0$  appears to conflict with the result of Theorem 2.7.1, which states that  $\rho(T, E)$  is always zero. The paradox presented here, however, is easily resolved. The apparent conflict results not from mathematical error but rather because a different concept of true score has been used in each case. Assuming that the propensity distributions were identical among (true) pullets and among (true) cockerels, the true score of a pullet defined as an expected observed score, as in Section 2.3, would be the conditional probability of a pullet being classified as a cockerel, that is,

$$0 \cdot p_{00}/p_{0\bullet} + 1 \cdot p_{01}/p_{0\bullet} = p_{01}/p_{0\bullet}.$$

In general this value is greater than zero, the value of the Platonic true score defined in this section. Similarly, in the first sense, the true score for cockerels is just  $p_{11}/p_{1\bullet}$ , the conditional probability of a cockerel being classified as a cockerel.

The definition of true score developed in Section 2.3 and the definition of error score constructed from it in Section 2.4 are based on an operationally related conception. The definition of true score developed in this section and the error scores constructed from this definition are based on a Platonic conception. In general the two concepts and definitions do not agree. There is no problem here providing we keep in mind that in this book, unless we specifically note the contrary, whenever we speak formally of true score, we are referring to the former conception and definition and not to the latter. We do not suggest that the conception of true score that we have adopted here (and which we shall enlarge somewhat in later chapters) is the only useful concept of true score. Clearly it is not, as the above example would suggest.

The mode of the propensity distribution is a still different kind of true score. A generalization of Sutcliffe's approach based on the mode would be useful, although it is clear that this would not lead to the simple kinds of results which we shall obtain using the expectation. A fourth true score of interest could be the median of the propensity distribution. This definition would have the property that a person would remain above or below the median regardless of any monotonic transformation of the score scale which, as we have already pointed out, is an admissible operation for ordinal measurement scales. However, no general theory of these true-score models is available.

The Platonic concept of true score is not one which should be or is likely to be completely neglected. Indeed, factor analytic theory was originally conceived in Platonic terms. We simply point out here that the operationally related definition of true score which we usually adopt is in some ways a very convenient one theoretically. Perhaps the example we have given and its demonstration of what the concept of true score shall mean for us will help even further to dispel the mystery which apparently has surrounded the concept of true score.

On the other hand, under certain circumstances it is quite possible to relate the operational and Platonic conceptions of true score. The two concepts are identical whenever the Platonic true scores are such that the Platonic error scores  $E_{g*} = X_{g*} - T_{g*}$  obey the assumptions derived in Theorem 2.7.1. In the chicken-sexing example this does not occur in the given scales. Nevertheless it might seem reasonable in some situations to presume that objects (chickens or persons) may have true values on some property (e.g., height) but that successive measurements are "by chance" disturbed either positively or negatively in a random and unbiased fashion. If this is the case, the operationally related definition will be consistent with the Platonic conception of true score. Furthermore, as we pointed out earlier and shall prove in Section 5.4, the operationally related definition of true score implies the conception of a true score as the (average) score that a person would attain on a test of infinite length (the von Mises conception). Thus we are able to provide a semantic interpretation of the concept of the true score.

At the risk of reinstating much of the confusion which this discussion has been designed to dispel, we note that if we again assume the conditional identity of propensity distributions, we may bring the two conceptions of true score into coincidence for the chicken-sexing example. This may be done simply by transforming the Platonic true scores from  $(0, 1)$  to  $(p_{01}/p_{0*}, p_{11}/p_{1*})$  and retaining the observed-score scale as given. The proposed transformation is clearly admissible, in the sense of Section 1.5, since the scale of measurement is nominal.

This kind of transformation of the true-score scale is subject to objection because the true-score scale may sometimes have a very straightforward relationship to the natural numbers which this transformation disturbs. For example, suppose that a certain state has a head tax of one horse and that

certain persons are exempt from taxation. If we ask some long-time resident how much head tax he paid last year, his response will be either zero or one horse. We may, of course, consider such responses to be subject to error and may think in terms of the corresponding Platonic and operationally related concepts of true and error scores. Indeed the discussion of the previous example applies to this example except with respect to the final statement. While we felt only mildly disturbed at assigning the score (or code) values  $p_{01}/p_0$ , and  $p_{11}/p_1$ , to the true pullet and cockerel classification, we should experience much greater discomfort at transforming the given true tax scale to this new scaling. Clearly a person paid either zero horses or one horse and certainly not  $p_{01}/p_0$  horses. The basic difference of the two cases is that in the first the scale of measurement is only nominal and hence any one-to-one transformation is admissible, whereas in the second case the scale of measurement is *absolute* and no transformation is admissible. In psychology, however, there is seldom any fundamental theoretical reason for considering a scale to have absolute properties, except, as in this example, when an actual counting process is involved.

Finally, with respect to the syntactic definition of true score we have adopted here, it should be evident that a person's true score will depend on the various kinds of conditions under which the measurements are taken. For example, of all the conditions which affect measurement, we might choose to control lighting conditions. Suppose we set up two lighting conditions, one called "good" lighting and the other "bad". Then, over repeated experimentally independent observations for each condition, a true score for each person will be definable, and presumably these true scores will differ for each person over the two conditions. Also, if a third condition which involves a random sampling of the first two conditions is considered, a third true score can be defined. In Chapter 7 the first two of these true scores will be called a *specific true score* and the third, a *generic true score*. (See Exercise 2.10 for a further extension.)

In contrast to Platonic errors, the errors of measurement of the classical test theory model are by definition always unbiased. If certain effects—for example, poor lighting, excessive noise, and the inherent properties of the particular measurement scale adopted—have a biasing effect on the observed scores, this bias is represented in the classical model by a change in the true score, as in the pullet-cockerel example. This is seen as follows.

If we denote a biased error by  $E'_{ga}$ , the basic equation can be written as

$$X_{ga} = T'_{ga} + E'_{ga},$$

where neither  $T'_{ga}$  nor  $E'_{ga}$  is consistent with the assumptions usually made about true and error scores; for example,  $\mathcal{E}E'_{ga} \neq 0$ . The practical worker sometimes complains that mental test theory does not deal adequately with the bias  $\mathcal{E}E'_{ga}$ . The fact is that if one knew the size of the bias he would quickly redefine the observed measurement, substituting  $X_{ga} - \mathcal{E}E'_{ga}$  for  $X_{ga}$ . It is only when the size of the bias is unknown that the problem becomes serious.

Faced with the impossible task of making a correction for an unknown bias, the mental test theorist defines a new error of measurement

$$E_{ga} \equiv E'_{ga} - \mathcal{E}E'_{ga}$$

and a new true score

$$T_{ga} \equiv T'_{ga} + \mathcal{E}E'_{ga}.$$

Now the basic equation

$$X_{ga} = T_{ga} + E_{ga}$$

holds and  $E_{ga}$  is an unbiased error of measurement. This result is achieved at the expense of incorporating the bias  $\mathcal{E}E'_{ga}$  into the redefined true score  $T_{ga}$ .

The examiner may complain that the test theorist's redefined  $T_{ga}$  is a "true score" in which he has no interest. If so, he must revise his measurement procedures; for if  $T_{ga}$  is not of interest to him, then his observed measurements  $X_{ga}$  cannot be of interest to him either. The reason is that  $X_{ga}$  is simply  $T_{ga}$  plus an undesirable fluctuation.

## 2.10 Experimental Independence

At this point it is necessary to discuss formally the concept of experimental independence which has been discussed informally and employed earlier.

**Definition 2.10.1.** Measurements  $g$  and  $h$  are said to be *experimentally independent* if, for every  $a \in \mathcal{P}$ , the joint distribution function  $F_{gh,a}(x_{ga}, x_{ha})$  factors into the product  $F_{ga}(x_{ga})F_{ha}(x_{ha})$  of the marginal distribution functions.

Alternatively  $g$  and  $h$  are experimentally independent if, given that  $X_{ha} = x_{ha}$ , the conditional distribution function  $F_{ga}(x_{ga} | X_{ha} = x_{ha})$  of  $X_{ga}$  is identically equal to the unconditional distribution function  $F_{ga}(x_{ga})$ . If we use the terms *measurement procedure* and *experiment* synonymously, then experimental independence might well be called *measurement independence*. Since we are dealing here with a fixed individual  $a$  and since therefore  $\tau_{ga}$  and  $\tau_{ha}$  are constants, we have

**Theorem 2.10.2.** The assumption of experimental independence,

$$F_{gh,a}(x_{ga}, x_{ha}) = F_{ga}(x_{ga})F_{ha}(x_{ha}),$$

is equivalent to the assumption that the error random variables  $E_{ga}$  and  $E_{ha}$  are statistically independent and that the true scores  $\tau_{ga}$  and  $\tau_{ha}$  for all  $a$  are in fact constants, i.e., they do not change between measurements.

Given successive experimentally independent observations, the propensity distribution may be interpreted as describing the long-run relative frequency of various sets of possible measurements thus obtained. Similarly the observed-score distribution  $F_g$  may be taken as descriptive of the long-run relative frequency of various sets of values for randomly selected persons and for arbi-

trary replications. Thus we may speak interchangeably of expectations, variances, and so forth, over experimentally independent observations or with respect to the propensity distribution. In truth, however, experimental independence over an unlimited number of measurements is not generally approximable in practical mental testing work, and hence it is clear that true scores and the distribution of error scores are not directly measurable, as are observed scores.

As noted earlier, the availability of a few (approximately) experimentally independent observations is the rule in mental testing situations. An obvious exception would be a sequence of problems in which some problems depend on the successful solution of previous ones. Whether or not experimental independence can be assumed depends in part on the class of measurements being considered. Suppose, for example, that an examinee suffers from some temporary and chance indisposition during testing. This indisposition will affect his response to each item. If generalization is desired only to instances in which the student is thus indisposed, experimental independence may well be a reasonable assumption; but if generalization is desired to the larger class of situations in which the examinee's present health is not fixed, then the item measurements cannot be considered to be experimentally independent. However, the total test score can be considered to be one observation relevant to the examinee's behavior in this larger class of situations. We shall consider some of the practical implications of this discussion in Chapter 6. In Chapter 24 we shall show that the concept of experimental independence is intimately connected with the concept of local independence, which is the basic assumption of latent trait theory.

## 2.11 Linear Experimental Independence

We may define a concept similar to but weaker than that of experimental independence.

**Definition 2.11.1.** Measurements  $X_{ga}$  and  $X_{ha}$  are said to be *linearly experimentally independent* if  $\mathcal{E}(X_{ga} | x_{ha}) = \mathcal{E}X_{ga}$  and  $\mathcal{E}(X_{ha} | x_{ga}) = \mathcal{E}X_{ha}$ , for all  $a \in \mathcal{P}$ . Such measurements are also referred to as *distinct measurements*.

This assumption states, in effect, that the taking of one measurement does not affect the first moment of the distribution of the second measurement, and hence that the two measurements are uncorrelated.

We note that the condition for Theorem 2.7.1 (iii-iv) can be weakened so that we have the following:

**Theorem 2.11.2.** If, in Theorem 2.7.1 (iv), the assumption that  $X_{ga}$  and  $X_{ha}$  are experimentally independent is replaced by the assumption that  $X_{ga}$  and  $X_{ha}$  are linearly experimentally independent, then the result (2.7.1d) still holds, i.e.,

$$\rho(E_{g*}, E_{h*}) = 0.$$

The proof of this theorem follows the methods used in Theorem 2.7.1 and is left as an exercise (2.2) for the reader.

The assumption of linear experimental independence implies that

$$\mathcal{E}(E_{ga} | E_{ha} = e_{ha}) = \mathcal{E}(E_{ga}), \quad (2.11.1)$$

which may be contrasted with an alternate statement of the assumption of experimental independence, that is,

$$F(e_{ga} | E_{ha} = e_{ha}) = F(e_{ga}). \quad (2.11.2)$$

From this it is clear that linear experimental independence is a much weaker assumption than experimental independence, for clearly (2.11.2) implies (2.11.1) but is not implied by it. In fact (2.11.2) implies that

$$\mathcal{E}(E_{ga}^r | E_{ha} = e_{ha}) = \mathcal{E}(E_{ga}^r)$$

for all  $r > 0$ .

Actually the assumption of linear experimental independence is *mathematically* stronger than the assumption  $\rho(E_{g*}, E_{h*}) = 0$  of the classical model, which we adopt in Chapter 3. Conceptually, however, it is simpler. Some further implications of this assumption are presented in Section 10.2. Thus the classical theory of mental test scores introduced axiomatically in this chapter, and to be developed in detail in the next four chapters, can be based upon the specification of an interval scale, an assumption of finite variances, and an assumption of linear experimental independence.

## 2.12 Replicate Measurements

In statistical writing the term *replications* is used to refer to successive observations which are independently and identically distributed. In test theory, correspondingly, this term is used to refer to measurements having the same true scores and independent and identically distributed errors of measurement. Thus stated, the concept of replication denotes an equivalence relationship between observations. Thus we state

**Definition 2.12.1.** Measurements  $X_g$  and  $X_{g'}$  are equivalent in  $\mathcal{P}$  if  $\tau_{ga} = \tau_{g'a}$  and  $F(E_{ga}) \equiv F(E_{g'a})$  for all  $a \in \mathcal{P}$ .

However, we do not require so strong an equivalence relationship between repeated measurements for making inferences using the classical model. The classical model deals only with first- and second-order moments and consequently the repeated measurements need only have the same true scores and linearly experimentally independent errors with equal variances, even though higher-order moments of the distribution of the repeated measurements may differ. Stronger models are required if higher-order moments of the distributions are to be studied (see Eq. 10.2.1). We shall employ *replication* as a generic

term to denote repeated observations which are effectively equivalent insofar as the model being used is concerned.

The distribution of the random variable  $X_{ga}$ , referred to as the propensity distribution, may be interpreted semantically as describing the relative frequency of various sets of  $x_{ga}$ -values over repeated (linearly) independent measurements  $g$  on a fixed person  $a$ . We shall refer to such measurements as replications and say that  $X_{ga}$  is defined *over replications* or *in a replications space*. Similarly the distribution of the random variable  $X_{g*}$  may be interpreted semantically as describing the relative frequency of values  $x_{g*}$  over repeated sampling of persons. Hence we shall say that  $X_{g*}$  is defined *over persons*. For the classical model, the replications need only be in terms of repeated linearly experimentally independent measurements having the same true scores and equal variances, and not necessarily of experimentally independent measurements. Such replications do not necessarily conform to replications from the propensity distribution except in regard to the first- and second-order moments, those quantities with which the classical model is concerned.

In using the classical model, which deals only with first- and second-order moments, expectations taken "over replications" are identical with expectations taken with respect to the propensity distribution. Thus, for example,

$$\mathcal{E}_k \sigma^2(E_{gak}) = \mathcal{E} \sigma^2(E_{ga}) = \sigma^2(E_{ga}),$$

which the reader is asked to prove in Exercise 2.8. For this reason, in the following chapter we are able to avoid reference to the propensity distribution and instead work directly with expectations taken over replications. We accomplish this by introducing the concept of parallel measurements.

### 2.13 Parallel Measurements and Parallel Forms

We have referred to successive measurements which can be considered equivalent with respect to the assumed model as replications. Because of the limited kinds of statement to be made using the classical model, it is not necessary to assume that repeated measurements have both identical true scores and identically distributed errors. We need only assume that they have identical true scores and linearly experimentally independent errors having equal variances. In the framework of the classical theory we refer to such measurements as *parallel measurements* and to the test forms from which they are obtained as *parallel tests*.

By  $\tau$ -equivalent measurements we shall mean those having the same true scores but (possibly) different error variances. Many of the standard test theory results can be obtained assuming only that repeated measurements are  $\tau$ -equivalent; it is not necessary to assume that they are parallel.

From this discussion and the derivations of this and the following chapters we shall see that, given the assumed finiteness of all variances, parallelism for repeated measurements, and linear experimental independence for distinct

measurements, the classical test theory model is a tautology rather than a model or a theory. By this we mean that the model must hold with respect to any given set of data. Thus, for most purposes, the empirical validity and utility of the model rest on the assumption of linear experimental independence and the availability of parallel measurements. The assumption of linear experimental independence is certainly a relatively weak assumption. In many testing applications it is one which we are generally prepared to make for at least a short series of observations. For estimation purposes, we usually require a minimum of only two parallel measurements, and in mental testing we are willing to suppose that we can obtain at least two approximately parallel measurements by methods discussed in Chapter 6.

Formally, then, we state

**Definition 2.13.1.** Distinct measurements  $X_{ga}$  and  $X_{g'a}$  are called *parallel measurements* if, for every subject  $a \in \mathcal{P}$ ,  $\tau_{ga} = \tau_{g'a}$ , and  $\sigma(E_{ga}) = \sigma(E_{g'a})$ .

Thus parallel measurements measure exactly the same thing in the same scale and, in a sense, measure it equally well for all persons. Some immediate consequences of this definition are given here. Other consequences are given in Section 3.3, where we discuss briefly the problem of verifying the parallelism of measurements.

We now state

**Theorem 2.13.2.** Distinct measurements  $X_{ga}$  and  $X_{g'a}$  are parallel measurements in  $\mathcal{P}$  if and only if in every subpopulation of  $\mathcal{P}$ ,

$$\mathcal{E}(X_{g*}) = \mathcal{E}(X_{g'*}), \quad (2.13.1)$$

*the expected values of parallel measurements are equal; and*

$$\sigma^2(X_{g*}) = \sigma^2(X_{g'*}), \quad (2.13.2)$$

*the observed-score variances of parallel measurements are equal.*

The proof of this theorem is left as an exercise (2.5) for the reader; the key to the proof is to consider all subpopulations of one person.

Two fundamental properties of parallel measurements are given in

**Theorem 2.13.3.** Let  $X_{1*}, X_{2*}, \dots$  be parallel measurements and  $Z$  be an arbitrary distinct measurement. Then

$$\rho(X_{1*}, X_{2*}) = \rho(X_{1*}, X_{3*}) = \dots = \rho(X_{2*}, X_{3*}) = \dots, \quad (2.13.3)$$

*all intercorrelations of parallel tests are equal; and*

$$\rho(X_{1*}, Z) = \rho(X_{2*}, Z) = \rho(X_{3*}, Z) = \dots, \quad (2.13.4)$$

*all parallel tests correlate equally with any other test.*

*Proof*

$$\rho(X_{1*}, X_{2*}) = \frac{\sigma(X_{1*}, X_{2*})}{\sigma(X_{1*})\sigma(X_{2*})} = \frac{\sigma(T_1 + E_1, T_2 + E_2)}{\sigma(X_1)\sigma(X_2)} = \frac{\sigma(T_1, T_2)}{\sigma(X_1)\sigma(X_2)}.$$

But  $T_1 \equiv T_2$ ; hence  $\sigma(T_1, T_2) = \sigma^2(T_1)$ , and

$$\rho(X_1, X_2) = \frac{\sigma^2(T_1)}{\sigma(X_1)\sigma(X_2)}.$$

Similarly

$$\rho(X_1, X_3) = \frac{\sigma^2(T_1)}{\sigma(X_1)\sigma(X_3)}.$$

But  $\sigma(X_2) = \sigma(X_3)$ , by parallelism.  $\square$

The advantage of our definition of parallel measurements in contrast with previous definitions is that *under our definition, if  $X_{ga}$  and  $X_{ha}$  are parallel in  $\mathcal{P}$ , then they are parallel in any subset of  $\mathcal{P}$* , a requirement of some intuitive and mathematical appeal. In previous treatments of test theory, measurements satisfying Eqs. (2.13.1) through (2.13.4) in  $\mathcal{P}$ , but not necessarily in every subpopulation of  $\mathcal{P}$ , were taken to be parallel. We shall call such measurements *pseudoparallel*. This weaker definition is adequate to derive the standard results of the classical model for the population  $\mathcal{P}$ . However, these results would not then necessarily be true for subpopulations of  $\mathcal{P}$ . (Also see Exercise 2.11.)

Our definition of parallel measurements is an explication of a general concept, an explicandum whose meaning has been rather imprecisely stated in previous writings. Usually this definition will be entirely satisfactory. However, it does have the weakness that if  $g$  and  $h$  are parallel in  $\mathcal{P}$ , they need not be parallel in some *larger* population. We presume henceforth that these conditions hold in the largest population we shall encounter in any study and hence refer to measurements simply as being parallel, dropping the technical qualifier, "in  $\mathcal{P}$ ". For an interesting discussion of the foundations of the classical model and the role of parallel measurements in this model the reader is referred to Tryon (1957). In Chapter 3 and following chapters, it will be convenient to adopt the following definition of parallel measurements:

**Definition 2.13.4.** Distinct measurements  $X_{g*}$  and  $X_{g'*}$  are parallel if

$$T_{g*} \equiv T_{g'*} \quad \text{and} \quad \sigma^2(X_{g*}) = \sigma^2(X_{g'*})$$

in every subpopulation of  $\mathcal{P}$ .

The practical problems involved in obtaining parallel measurements will be discussed in some detail in Section 6.4.

Again, explicitly we state

**Definition 2.13.5.** Distinct measurements  $X_{ga}$  and  $X_{ha}$  are  $\tau$ -equivalent if, for all  $a$ ,  $\tau_{ga} = \tau_{ha}$ .

Then we have

**Theorem 2.13.6.** Distinct measurements  $X_{ga}$  and  $X_{ha}$  are  $\tau$ -equivalent if and only if

$$\mathcal{E}(X_{g*}) = \mathcal{E}(X_{h*})$$

for every subpopulation of  $\mathcal{P}$ .

Thus, equivalently, we have

**Definition 2.13.7.** Distinct measurements  $X_{g*}$  and  $X_{g'*}$  are  $\tau$ -equivalent if  $\mathcal{E}(X_{g*}) = \mathcal{E}(X_{g'*})$  in every nonnull subset of  $\mathcal{P}$ .

Finally we state

**Definition 2.13.8.** Measurements  $X_{g*}$  and  $X_{h*}$  are *essentially  $\tau$ -equivalent* if they are distinct and if, for all  $a$ ,  $\tau_{ga} = a_{gh} + \tau_{ha}$ , where  $a_{gh}$  is a constant.

Chapter 10 will be devoted to the theory of  $\tau$ -equivalent measurements and their use in estimating higher-order moments of the true- and error-score distributions. The concept of essential  $\tau$ -equivalence plays a central role in Chapter 4, in which it is shown that two important test theoretic indices are equal if and only if all items of a test are essentially  $\tau$ -equivalent.

The  $\tau$ -equivalence and essential  $\tau$ -equivalence relationships are *equivalence relationships* (in the technical sense of that phrase) which are less strong than the parallelism relationship. The  $\tau$ -equivalence relationship implies that although the measurements have the same true scores for all persons, these true scores are not necessarily measured equally well for each person by the two tests. Further, essential  $\tau$ -equivalence introduces the possibility that the true scores on the two measurements may differ by as much as an additive constant.

## Exercises

- 2.1. Prove part (iii) of Theorem 2.7.1.
- 2.2. Prove Theorem 2.11.2 by first proving that  $\mathcal{E}(E_{g*} | e_{h*}) = 0$  for all possible values of  $E_{h*}$ .
- 2.3. For the example at the end of Section 2.5 verify that  $\mathcal{E}(T_{g*}) = \mathcal{E}(X_{g*})$ ,  $\sigma^2(X_{g*}) = \mathcal{E}_a\sigma^2(X_{ga}) + \sigma^2(\mathcal{E}X_{ga})$ ,  $\mathcal{E}(E_{g*}) = 0$ ,  $\sigma^2(E_{g*}) = \mathcal{E}_a\sigma^2(E_{ga})$ , and  $\rho(E_{g*}, T_{g*}) = 0$ .
- 2.4. Verify these same results for the pullet-cockerel example of Section 2.9.
- 2.5. Prove Theorem 2.13.2.
- 2.6. Assuming only that  $\sigma^2(X_{g*}) < \infty$  and  $X_{g*} = T_{g*} + E_{g*}$ , show by counterexample that both  $\sigma^2(T_{g*})$  and  $\sigma^2(E_{g*})$  are not necessarily finite.
- 2.7. Assume now that  $X_{g*}$ ,  $T_{g*}$ , and  $E_{g*}$  are the random variables constructed in Sections 2.2 through 2.6. Show that  $\sigma^2(T_{g*}) < \infty$  and  $\sigma^2(E_{g*}) < \infty$ .
- 2.8. Show that

$$\mathcal{E}_k\sigma^2(E_{gak}) = \mathcal{E}\sigma^2(E_{ga}) = \sigma^2(E_{ga}).$$

- 2.9. Prove Theorem 2.6.2, paying particular attention to the case where  $\sigma^2(X)$  is infinite.
- 2.10. Suppose that a verbal and a quantitative test are available and that the use of each under specified conditions gives rise to measurements  $X$  and  $Y$ , respectively. Show that if a fair coin is tossed to determine which test each examinee shall take, a new random variable can be defined over examinees for the measurements thus generated and that corresponding true and error scores can be defined to satisfy the usual definitions of true and error score.
- 2.11. Suppose that measurements  $X$  and  $Y_1$  are pseudoparallel and  $Y_1$  and  $Z$  are pseudoparallel. Show that  $X$ ,  $Y$ , and  $Z$  need not form a pseudoparallel set.
- 2.12. Suppose  $X_g$  and  $X_h$  are not distinct. Given  $X_g$ , show how to construct a residual measurement  $X_h^*$  of  $X_h$ , such that  $X_g$  and  $X_h^*$  are distinct, and no information concerning first- and second-order moments of  $X_g$  and  $X_h$  is lost and that no information concerning  $T_g$  and  $T_h$  is lost.
- 2.13. Let  $X_i = T_i + E$ ,  $i = 1, 2, 3, 4$ , be distinct measurements with

$$T_i = \sum_{j=1} a_{ij} \theta_j.$$

Show that the  $a_{ij}$  can be determined so that

$$\begin{aligned}\mathcal{E}(X_1) &= \mathcal{E}(X_2) = \mathcal{E}(X_3), \\ \sigma(X_1) &= \sigma(X_2) = \sigma(X_3), \\ \rho(X_1, X_2) &= \rho(X_1, X_3) = \rho(X_2, X_3), \\ \rho(X_1, X_4) &= \rho(X_2, X_4) = \rho(X_3, X_4),\end{aligned}$$

but that  $X_1$ ,  $X_2$ , and  $X_3$  are not parallel.

- 2.14. Discuss as completely as possible the advantages and disadvantages of basing a true-score theory on the median of the propensity distribution.
- 2.15. What major difficulty is immediately encountered in attempting to define *true score* as the mode of the propensity distribution?
- 2.16. For arbitrary random variables  $X$ ,  $Y$ , and  $Z$ , show that

$$\sigma(X, Y) = \mathcal{E}\sigma(X, Y | z) + \sigma[\mathcal{E}(X), \mathcal{E}(Y) | z].$$

- 2.17. Suppose that the observed score  $X_a$  of each person follows a normal distribution with expected value  $\tau$  and variance 1, and that the distribution of  $\tau$  values over people is normal with mean  $\mu$  and variance 1. Let  $T$  denote the true-score variable, taking values  $\tau$ .
- For each person  $a$  let the error random variable be defined by  $E_a = X_a - \tau_a$ . Show that  $E_a$  is normally distributed with mean 0 and variance 1.
  - Let  $E$  be the error random variable when the person is randomly selected. Show that  $E$  is normally distributed with mean 0 and variance 1.
  - Let  $X$  be the observed-score random variable for a randomly selected person. Show that  $X$  is normally distributed with expected value  $\mu$  and variance 2, and that the  $T$  and  $E$  are uncorrelated in every nonnull subpopulation.

**References and Selected Readings**

- CARNAP, R., Logical foundations of probability. Chicago: University of Chicago Press, 1950. Second edition, 1962.
- CRONBACH, L. J., and GOLDINE C. GLESER, Psychological tests and personnel decisions, 2nd ed. Urbana: University of Illinois Press, 1965.
- FREEMAN, H., *Introduction to statistical inference*. Reading, Mass.: Addison-Wesley, 1963.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.
- GUTTMAN, L., A basis for analyzing test retest reliability. *Psychometrika*, 1945, **10**, 255-282.
- GUTTMAN, L., A special review of Harold Gulliksen, Theory of mental tests. *Psychometrika*, 1953, **18**, 123-130. (a)
- GUTTMAN, L., Reliability formulas that do not assume experimental independence. *Psychometrika*, 1953, **18**, 225-239. (b)
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*, Vol. I. New York: Hafner, 1958.
- LAZARSFELD, P. F., Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science*, Vol. 3. New York: McGraw-Hill, 1959.
- LINDLEY, D. V., *Introduction to probability and statistics from a Bayesian point of view. Part I: Probability*. Cambridge: Cambridge University Press, 1965.
- LOEVINGER, JANE, Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, **3**, 635-694.
- NOVICK, M. R., The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, **3**, 1-18.
- SUTCLIFFE, J. P., A probability model for errors of classification. I: General considerations. *Psychometrika*, 1965, **30**, 73-96.
- THORNDIKE, R. L., Reliability. In *Proceedings of the 1963 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1964, 23-32.
- TRYON, R. C., Reliability and behavior domain validity: reformulation and historical critique. *Psychological Bulletin*, 1957, **54**, 229-249.

**Part 2**

**THE CLASSICAL  
TEST THEORY MODEL**



## CHAPTER 3

# BASIC EQUATIONS OF THE CLASSICAL MODEL FOR TESTS OF FIXED LENGTH

### 3.1 The Classical Linear Model: Restatement of Assumptions

In this chapter, we shall deal entirely with formal manipulations within the classical test theory model in order to deduce relationships among various quantities of interest. Since observed scores are directly observable and true scores and error scores are not, we shall express the various parameters of true and error scores in terms of parameters of observed scores. We have already presented one result in these terms, namely, that the expected true score is equal to the expected observed score. Generally this kind of result requires the use of parallel measurements. In this section, we present the assumptions of the model and in the following two sections we shall derive the very basic equations of the model. Interpretations and applications of these equations and the axioms of the model are discussed in Sections 3.4, 3.5, and 3.6. The remaining sections of the chapter treat various regression theory results for the classical model.

We caution the reader at the outset that the formulas developed in this chapter are not in every case the best ones available for estimating parameters of the latent variable distributions whenever only limited sample information about the observed-score distribution is given. Only when samples are very large is it possible to obtain good estimates of the parameters of the latent variable distributions by substituting sample analogs of observed-score distribution parameters directly into these formulas. Nevertheless the formulas developed in this chapter are necessary for an understanding of the classical model and as a basis for the development of the estimation procedures contained in Chapters 7 through 9.

So that this chapter may be read without reference to the more technical parts of Chapter 2 we begin with a formal restatement of the assumptions of the classical model as derived in the previous chapter. Let  $X$  be a random variable defined over a population  $\mathcal{P}$  of persons and taking values  $x$  corresponding to the observed values obtained on different persons. Let  $T$  be the associated true-score random variable taking values  $\tau$  corresponding to the

unobserved true scores of these people.\* For a fixed person, the true score is constant, but the observed score and error score are random variables. The error random variable  $E$  is defined by the linear relation

$$X = T + E \quad (3.1.1)$$

over  $\mathcal{P}$ . The (unobserved) error score  $e$  is just the difference between the observed value  $x$  and the true score  $\tau$ . Let

$$\mathcal{E}(X) \equiv \mathcal{E}_a(X) \equiv \mu_X,$$

where (here and throughout Chapters 3 through 6) the expectation sign with or without the subscript  $a$  indicates expectation over persons, and let us denote the variances of  $X$ ,  $T$ , and  $E$  by  $\sigma_X^2$ ,  $\sigma_T^2$ , and  $\sigma_E^2$ , respectively. When subscripts are necessary to distinguish different measurements we shall denote the corresponding variances by  $\sigma^2(X_1)$ ,  $\sigma^2(T_1)$ ,  $\sigma^2(E_1)$ , and so on to avoid the use of secondary subscripts. We shall denote the correlations between  $X$  and  $E$ ,  $T$  and  $E$ , and  $E_1$  and  $E_2$ , where  $E_1$  and  $E_2$  are the error random variables for two distinct measurements, by  $\rho_{XE}$ ,  $\rho_{TE}$  and  $\rho(E_1, E_2)$ , respectively. We shall denote the corresponding covariances by  $\sigma_{XE}$ ,  $\sigma_{TE}$  and  $\sigma(E_1, E_2)$ . Similarly we may write  $\sigma(X) \equiv \sigma_X$ ,  $\sigma(X, Y) \equiv \sigma_{XY}$ ,  $\mu(X) \equiv \mu_X$ , and  $\rho(X, Y) \equiv \rho_{XY}$ . We shall write  $\beta(X | y) \equiv \beta_{XY}$  for the regression coefficient of  $X$  on  $y$ .

The assumptions of the classical model, which we shall adopt throughout Chapters 3, 4, and 5, are represented by Eq. (3.1.1) and the following equations:

$$\mathcal{E}E = 0, \quad (3.1.2)$$

$$\rho_{TE} = 0, \quad (3.1.3)$$

and

$$\rho(E_1, E_2) = 0, \quad (3.1.4)$$

$$\rho(E_1, T_2) = 0. \quad (3.1.5)$$

which are taken to hold in every nonnull subpopulation of  $\mathcal{P}$ . Again to avoid triviality, we assume that

$$0 < \sigma_X^2 < \infty, \quad 0 < \sigma_T^2 < \infty, \quad \text{and} \quad 0 < \sigma_E^2 < \infty. \quad (3.1.6)$$

### 3.2 Expectations, Variances, and Correlations

From these assumptions we again immediately obtain the result

$$\mathcal{E}T = \mathcal{E}X \equiv \mu_X = \mu_T, \quad (3.2.1)$$

say, for

$$\mathcal{E}T = \mathcal{E}(X - E) = \mathcal{E}X - \mathcal{E}E = \mathcal{E}X.$$

*The expected true score is equal to the expected observed score.*

---

\* The reader should now review the *Notation* section, noting in particular that the symbol  $T$  is the upper-case Greek “tau”.

Also, since  $\mathcal{E}E = 0$  in every nonnull subpopulation of  $\mathcal{P}$ , in particular the expectation is zero in the subpopulation of persons with any specified true score  $\tau$ ; that is,

$$\mathcal{E}(E | \tau) = 0, \quad \text{for all } \tau. \quad (3.2.2)$$

Then

$$\mathcal{E}(X | \tau) = \mathcal{E}(T + E | \tau) = \mathcal{E}(T | \tau) + \mathcal{E}(E | \tau),$$

or

$$\mathcal{E}(X | \tau) = \tau. \quad (3.2.3)$$

*The regression of observed score on true score is a straight line through the origin with unit slope.*

From (3.1.1) we have

$$\sigma_X^2 = \sigma^2(T + E) = \sigma_T^2 + \sigma_E^2 + 2\sigma_{TE}.$$

However,  $\sigma_{TE} = \rho_{TE}\sigma_T\sigma_E = 0$ , since  $\rho_{TE} = 0$ , and hence

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \quad (3.2.4)$$

*The observed-score variance in the population of persons is equal to the sum of the true-score variance and the error-score variance.*

By definition,

$$\begin{aligned} \sigma(X, T) &= \mathcal{E}(XT) - \mathcal{E}(X)\mathcal{E}(T) = \mathcal{E}[(T + E)T] - \mathcal{E}(T + E)\mathcal{E}(T) \\ &= \mathcal{E}T^2 + \mathcal{E}(ET) - (\mathcal{E}T)^2 - (\mathcal{E}E)\mathcal{E}(T) = \sigma_T^2 + \mathcal{E}(ET) - \mathcal{E}(E)\mathcal{E}(T). \end{aligned}$$

But  $\mathcal{E}E = 0$ , and since  $\sigma_{TE} = 0$ , we have  $\mathcal{E}ET = 0$ . Hence

$$\sigma_{XT} = \sigma_T^2.$$

But  $\rho_{XT}^2 = \sigma_{XT}^2/\sigma_X^2\sigma_T^2$ , and hence

$$\rho_{XT}^2 = \sigma_T^2/\sigma_X^2. \quad (3.2.5)$$

*The square of the correlation between observed scores and true scores equals the ratio of the true-score variance to the observed-score variance.*

This identification of the coefficient of determination  $\rho_{XT}^2$  with the variance ratio  $\sigma_T^2/\sigma_X^2$  is a result of the linearity of the regression shown in Eq. (3.2.3). As we have shown in the preceding chapter, however, this linearity results from the way in which we have constructed our true- and error-score random variables. For this reason, we shall often be able, throughout the next several chapters, to study correlation coefficients by examining the variance ratios with which they are identified and thus reduce such correlation studies to problems in variance analysis.

From (3.2.4) and (3.2.5) we obtain

$$\rho_{XT}^2 = 1 - (\sigma_E^2/\sigma_X^2). \quad (3.2.6)$$

The result

$$\rho_{XE}^2 = \sigma_E^2/\sigma_X^2 \quad (3.2.7)$$

may be derived by a method similar to that followed in obtaining Eq. (3.2.5); it is left as an exercise (3.8) for the reader. The following relation is now obvious:

$$\rho_{XT}^2 + \rho_{XE}^2 = 1. \quad (3.2.8)$$

### 3.3 Relationships Based on Parallel Measurements

Now let us assume that we have measurements  $X$  and  $X'$  such that

$$X = T + E \quad \text{and} \quad X' = T + E' \quad (3.3.1a)$$

and

$$\sigma^2(E) = \sigma^2(E') \quad (3.3.1b)$$

in every nonnull subpopulation of  $\mathcal{P}$ . Such measurements are called *parallel measurements*. Two immediate implications of this assumption are that

$$\mathcal{E}X = \mathcal{E}X' \quad (3.3.2)$$

and

$$\sigma^2(X) = \sigma^2(X') \quad (3.3.3)$$

in every subpopulation of  $\mathcal{P}$ . A more complete discussion of parallel measurements has been given in Section 2.13.

The correlation between any pair of parallel measurements may now be easily obtained. Since correlation coefficients are not affected by change of origin or scale of measurement, we may assume, with no loss of generality, that  $\mathcal{E}X = \mathcal{E}X' = \mathcal{E}T = 0$  and obtain

$$\begin{aligned} \rho(X, X') &= \frac{\sigma(X, X')}{\sigma(X)\sigma(X')} = \frac{\mathcal{E}XX'}{\sigma(X)\sigma(X')} = \frac{\mathcal{E}(T+E)(T+E')}{\sigma(X)\sigma(X')} \\ &= \frac{\mathcal{E}T^2 + \mathcal{E}TE + \mathcal{E}TE' + \mathcal{E}EE'}{\sigma(X)\sigma(X')} . \end{aligned} \quad (3.3.4)$$

But  $\mathcal{E}TE = \mathcal{E}TE' = \mathcal{E}EE' = 0$  by Eqs. (3.1.5) and (3.1.4), and  $\sigma(X) = \sigma(X') = \sigma_X$ ; hence

$$\rho(X, X') = \sigma_T^2 / \sigma_X^2 \geq 0. \quad (3.3.5)$$

When considering parallel measurements  $X$  and  $X'$  we shall, for simplicity of notation, usually write

$$\rho_{XX'} \equiv \rho(X, X'). \quad (3.3.5a)$$

Comparing (3.2.5) with (3.3.5) and (3.3.5a) we obtain

$$\rho_{XT}^2 = \rho_{XX'}. \quad (3.3.5b)$$

*The square of the correlation between observed scores and true scores is equal to the correlation between parallel measurements.* Thus, assuming that at least one pair of parallel measurements can be obtained, we have succeeded in expressing

an unobservable quantity  $\rho_{XT}^2$  in terms of  $\rho_{XX'}$ , a parameter of a (bivariate) observed-score distribution. Hence the estimation of  $\rho_{XT}^2$  is reduced to the estimation of  $\rho_{XX'}$ .

From (3.3.5) it is evident that

$$\sigma_T^2 = \sigma_{XX'}. \quad (3.3.5c)$$

*The true-score variance (an unobservable quantity) is equal to the covariance between parallel measurements (a potentially observable quantity).* From this result it is easy to obtain

**Theorem 3.3.1.** Let  $X_1, X_2, X_3, \dots$  be parallel measurements and  $Z$  an arbitrary distinct measurement. Then

$$\rho(X_1, X_2) = \rho(X_1, X_3) = \dots = \rho(X_2, X_3) = \dots \quad (3.3.6a)$$

and

$$\rho(X_1, Z) = \rho(X_2, Z) = \rho(X_3, Z) = \dots. \quad (3.3.6b)$$

*All intercorrelations of parallel tests are equal and all parallel tests correlate equally with any other test.* (See Exercises 3.2 and 3.3 and Section 2.13.)

Also from (3.3.5) we have

$$\sigma_T^2 = \sigma_X^2 \rho_{XX'}, \quad (3.3.7)$$

which expresses the true-score variance again in terms of potential observables: *The true-score variance is equal to the product of the observed-score variance and the correlation between parallel measurements.* Combining (3.2.4) and (3.3.7) we have  $\sigma_E^2 = \sigma_X^2 \rho_{XX'} + \sigma_E^2$ , or

$$\sigma_E^2 = \sigma_X^2 (1 - \rho_{XX'}). \quad (3.3.8)$$

*The error-score variance is equal to the product of the observed-score variance and one minus the correlation between parallel measurements.* Equation (3.3.8) may be written

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}, \quad (3.3.9)$$

which again expresses an unobservable in terms of potential observables. Finally, from (3.2.7), using (3.2.4) and (3.3.5), we have

$$\rho_{XE} = \sqrt{\sigma_E^2 / \sigma_X^2} = \sqrt{1 - (\sigma_T^2 / \sigma_X^2)} = \sqrt{1 - \rho_{XX'}}. \quad (3.3.10)$$

The process of determining whether two or more tests are sufficiently parallel for the purposes at hand begins with the verification of the correctness of Eqs. (3.3.2), (3.3.3), (3.3.6a), and (3.3.6b) for all tests under consideration and for the population at hand. The validity of these equations for the population in question is a necessary, but not a sufficient, condition of parallelism. On the other hand, the validity of these equations for the population at hand is sufficient to establish the applicability of the standard results to this population.

In practice, the inference of (rigorous) parallelism is based not only on formal statistical procedures but also on a content analysis of the tests. Votaw has contributed one formal procedure for the statistical evaluation of Eqs. (2.13.1), (2.13.2), and (3.3.6a) which is given by Gulliksen (1950, pp. 187–192). Also see Section 10.5. More broadly, Chapters 8 through 10 detail the kinds of inference that can be made when the (rigorous) parallelism of repeated measurements is subject to question.

### 3.4 Definitions, Interpretations, and Applications

Each of the quantities identified in the previous sections has an important test theoretical use and most have descriptive names by which they may be identified. The quantity  $\mu_X$ , which is equal to both  $\mathcal{E}X$  and  $\mathcal{E}T$ , is often referred to as the *mean observed score* or the *mean true score*. The variance of the error scores  $\sigma_E^2 \equiv \sigma^2(E_{g*})$  is referred to as the variance of the errors of measurement or simply as the *error variance*, and its square root  $\sigma_E$  is referred to as the standard deviation of the errors of measurement or simply as the *standard error of measurement*. To differentiate  $\sigma(E_{g*})$  from  $\sigma(E_{ga})$ , where  $E_{ga}$  is the error random variable for a fixed person  $a$ , the former is often called the *average standard error of measurement in  $\mathcal{P}$* , and the latter, the *standard error of measurement for fixed person  $a$*  (see Section 2.4).

That  $\sigma_E^2$  may be taken as a measure of the average imprecision of a measurement may easily be demonstrated. From Chebychev's inequality (which holds for any distribution having finite variance), we have

$$\text{Prob } [|X_{ga} - \tau_{ga}| \geq k\sigma(X_{ga})] \leq 1/k^2 \quad (3.4.1)$$

for all  $k > 1$ . Since the length of the bound is directly in proportion to  $\sigma(X_{ga})$  and hence (see Chapter 2) to  $\sigma(E_{ga})$ , we may consider  $\sigma^2(E_{ga})$  to be a measure of the imprecision of the measurement  $X_{ga}$  relative to the true score  $\tau_{ga}$ . Generally, since  $\sigma^2(E_{g*}) = \mathcal{E}_a \sigma^2(E_{ga})$ , the former may be considered to be a measure of the imprecision of the measurement  $X_g \equiv X_{g*}$  relative to the true-score random variable  $T_g$  in the population  $\mathcal{P}$ . It is necessary, however, to point out that since  $\sigma_E^2$  varies with the scale of measurement, that is, with  $\sigma_X^2$ , it cannot be interpreted in an absolute way as a measure of imprecision unless one presumes that the given scale of measurement is absolute. This point will be made in more detail in Section 7.4.

The Chebychev inequality may be used to obtain confidence interval statements about examinees' true scores. Suppose, for example, that  $\sigma^2(E_{g*}) = 10$  and that we are willing to assume, as a first approximation, that the error variances for all people are equal. Then, given that a person's score is  $x_{ga} = 110$  and taking  $k = 2$  in Eq. (3.4.1), we have  $|\tau_{ga} - 110| \leq 20$  with confidence 0.75. In practice this is unlikely to be a useful approach since these bounds are typically very broad and better methods are often available (see Section 3.7).

The *reliability of a test* is defined as the squared correlation  $\rho_{XT}^2$  between observed score and true score. From the relation  $\rho_{XT}^2 = \sigma_T^2/\sigma_X^2$  we see that *the reliability of a test is a measure of the degree of true-score variation relative to observed-score variation*. From the relation

$$\rho_{XT}^2 = \sigma_T^2/\sigma_X^2 = 1 - (\sigma_E^2/\sigma_X^2),$$

we see that if the variance of the errors of measurement is zero, then the reliability of the test will be unity, while if the variance of the errors of measurement is equal to the observed-score variance, then the true-score variance  $\sigma_T^2$  must be zero and the reliability of the test must be zero. Since  $\sigma_X^2 \geq \sigma_E^2$  (the inequality is strict if we assume  $\sigma_E^2 > 0$ ), it follows that *reliability is a number in the interval from 0 to 1*. From this relation we also see that *reliability is an inverse measure of error variance relative to observed variance*. From (3.2.3) we see that the regression of observed scores on true scores is linear. The reliability coefficient is a measure of the strength of this linear relationship.

Since

$$\rho_{XT}^2 = \sigma_T^2/\sigma_X^2 = \rho_{XX'},$$

where  $X$  and  $X'$  are parallel measurements, any one of these quantities might be taken as the definition of reliability. We choose the first of these quantities to emphasize the fact that reliability may be *defined* without using the concept of parallel measurements and because this quantity proves useful in models of wider application. The quantity  $\rho_{XT} = \sigma_T/\sigma_X$  is sometimes referred to as the *index of reliability*. The quantities  $\sigma_E^2$  and  $\rho_{XT}^2$  are important measures of imprecision and precision of tests. These and related quantities will be a major subject of study throughout Part II of this book.

### 3.5 The Validities of a Test

The most meaningful measures of the value of a test are its validities. We assess validity according to the degree of relationship of a test with another variable of interest. In many situations it is convenient to restrict attention to linear relationship and then the correlation coefficient is the appropriate measure of relationship. Thus we state the technical

**Definition 3.5.1.** The *validity coefficient* of a measurement  $X$  with respect to a second measurement  $Y$  is defined as the absolute value of the correlation coefficient

$$\rho_{XY} = \sigma_{XY}/\sigma_X\sigma_Y. \quad (3.5.1)$$

It is evident that this definition is symmetric, i.e., the validity coefficient of one measurement with respect to a second measurement is equal to the validity coefficient of the second measurement with respect to the first. The validity coefficient of a measurement can therefore be stated only in relation to a second measurement; thus it makes no sense to speak of *the* validity coefficient of a

measurement. When speaking of the validity coefficient of a measurement, we shall always make explicit the second measurement unless the identity of this measurement can be clearly inferred from the context.

In the classical model, the validity coefficient is the usual explication of the general concept of validity. For this model, therefore, it will be convenient to use the simpler term *validity* in place of the more technically precise term *validity coefficient*.

For convenience we may here summarize some results on covariances and correlations which we shall use later in the study of validity theory. A simple technique for evaluating covariances is based upon the following result of probability theory:

$$\sigma(aU + bV, cX + dY) = ac\sigma_{UX} + ad\sigma_{UY} + bc\sigma_{VX} + bd\sigma_{VY}, \quad (3.5.2)$$

where  $U$ ,  $V$ ,  $X$ , and  $Y$  are random variables and  $a$ ,  $b$ ,  $c$ , and  $d$  are constants. For example, if  $U \equiv X$ , say, and  $V \equiv Y$ , we have

$$\sigma(aX + bY, cX + dY) = ac\sigma_X^2 + (ad + bc)\sigma(X, Y) + bd\sigma_Y^2. \quad (3.5.2a)$$

If  $X$  and  $X'$  are parallel measurements, a pertinent further example is

$$\begin{aligned} \sigma(X, X') &= \sigma[(T + E_1)(T + E_2)] \\ &= \sigma_T^2 + \sigma(TE_2) + \sigma(TE_1) + \sigma(E_1, E_2) = \sigma_T^2. \end{aligned} \quad (3.5.3)$$

Now let us consider two possibly nonparallel measurements  $X_1$  and  $X_2$ . First we note that

$$\begin{aligned} \sigma(X_1, X_2) &= \sigma[(T_1 + E_1), (T_2 + E_2)] \\ &= \sigma(T_1, T_2) + \sigma(T_1, E_2) + \sigma(T_2, E_1) + \sigma(E_1, E_2) \\ &= \sigma(T_1, T_2). \end{aligned} \quad (3.5.4)$$

The covariance between observed scores is equal to the covariance between true scores. If  $X_1$  and  $X_2$  are parallel measurements, then  $T_1 \equiv T_2$ , and hence

$$\sigma(T_1, T_2) = \sigma_T^2. \quad (3.5.5)$$

In a similar manner it is easily shown that, for possibly nonparallel measurements  $X_1$  and  $X_2$ ,

$$\sigma(T_1, X_2) = \sigma(X_1, T_2) = \sigma(T_1, T_2), \quad (3.5.6)$$

and that, for example,

$$\sigma(X_1, E_1) = \sigma^2(E_1). \quad (3.5.7)$$

The reader can demonstrate these relationships, which are given as Exercises 3.4 and 3.5.

We may also note that (3.5.5) holds even if the assumption of parallelism is relaxed, so that, possibly,  $\sigma^2(E_1) \neq \sigma^2(E_2)$  (see Exercise 3.6). Measurements satisfying (3.3.1) but not (3.3.1a) are said to be  $\tau$ -equivalent. For  $\tau$ -equivalent measurements  $X_1$ ,  $X_2$ , and  $X_3$  the reader can easily show that

$$\sigma(X_1, X_2) = \sigma(X_1, X_3) = \sigma(X_2, X_3), \quad (3.5.8)$$

although the corresponding relation involving correlations does not hold (see Exercise 3.7).

From (3.5.4) we have

$$\rho(X_1, X_2) = \frac{\sigma(T_1, T_2)}{\sigma(X_1)\sigma(X_2)}. \quad (3.5.9)$$

If  $X_1$  and  $X_2$  are parallel measurements, then  $\sigma(T_1, T_2) = \sigma_T^2$  and  $\sigma^2(X_1) = \sigma^2(X_2) = \sigma_X^2$ , say, and (3.5.9) reduces to (3.3.5), the reliability of the test. Thus *the reliability of a test is just its validity with respect to a parallel test*, as we have previously seen in (3.3.5b).

### 3.6 An Alternative Statement of the Classical Model\*

Our initial choice of assumptions was based on our direct derivation of them from more basic principles. It is certainly true, however, that the classical model can be stated in other ways by taking the linear model, Eq. (3.1.1), and selecting axioms which, in total, are equivalent to those given in Section 3.1. One such set of axioms has been given by Koutsopoulos (1964). Although the Koutsopoulos axioms may not have quite the “axiomatic flavor” of the statements in Section 3.1, they are of considerable interest. We shall restate these axioms here in a form consistent with our approach and then show that they are, thus stated, equivalent to those of Section 3.1. Paraphrasing the Koutsopoulos axioms, we have

*Alternative assumptions for the classical model:* In every nonnull subpopulation

$$(i) \quad \mathcal{E}X = \mathcal{E}T = \mu_X, \quad (3.6.1)$$

say;

$$(ii) \quad \mathcal{E}(X_1T_2) = \mathcal{E}(T_1T_2); \quad (3.6.2)$$

and, if  $X_1$  and  $X_2$  are distinct measurements, then

$$(iii) \quad \mathcal{E}(X_1X_2) = \mathcal{E}(X_1T_2). \quad (3.6.3)$$

First we show that these axioms imply those of Section 3.1.

Defining  $E \equiv X - T$ , and using (3.6.1), we have

$$\mathcal{E}E = \mathcal{E}(X - T) = 0,$$

---

\* Reading of this section may be omitted without loss of continuity.

thus establishing the axiom (3.1.2). Now

$$\begin{aligned}\rho_{TE} &= \frac{\sigma(T, E)}{\sigma_T \sigma_E} = \frac{\sigma[T, (X - T)]}{\sigma_T \sigma_E} \\ &= \frac{\mathcal{E}T(X - T) - (\mathcal{E}T)[\mathcal{E}(X - T)]}{\sigma_T \sigma_E} = \frac{\mathcal{E}XT - \mathcal{E}T^2}{\sigma_T \sigma_E}.\end{aligned}$$

Taking  $T_1 \equiv T_2 \equiv T$  in (3.6.2) shows that  $\rho(T, E) = 0$ , thus establishing (3.1.3). Also,  $\mathcal{E}X_1 X_2 = \mathcal{E}X_1 T_2 = \mathcal{E}T_1 T_2$ ; and hence, using both assumptions (3.6.2) and (3.6.3), we have

$$\begin{aligned}\rho(E_1, E_2) &= \frac{\sigma(X_1 - T_1, X_2 - T_2)}{\sigma(E_1)\sigma(E_2)} \\ &= \frac{\mathcal{E}T_1 T_2 - \mathcal{E}T_1 T_2 - \mathcal{E}T_2 T_1 + \mathcal{E}T_1 T_2}{\sigma(E_1)\sigma(E_2)} = 0,\end{aligned}$$

thus establishing the axiom (3.1.4). The axiom (3.1.5) may be derived in a similar manner. The statements (3.6.2) and (3.6.3) can easily be obtained from the axioms (3.1.2), (3.1.3), and (3.1.4) by the methods exhibited in Section 3.3, thus establishing the equivalence of the two sets of axioms. Certainly a fertile imagination could construct yet other sets of axioms equivalent to those stated in Section 3.2. While the Koutsopoulos axioms may lack the simplicity usually desired of axioms, they are convenient *mathematically*, and they do exhibit clearly the concept of covariation for linear models. This concept will find further explication in the Woodbury axiomatization which we shall present in Chapter 5.

### 3.7 Regression Theory for the Classical Model

For two arbitrary random variables  $X$  and  $T$ , the function  $\mathcal{E}(X | \tau)$ , the conditional expected value of  $X$  given that  $T = \tau$ , is called the *regression function of  $X$  on  $\tau$* . For the classical test theory model,  $\mathcal{E}(X | \tau) = \tau$  (Eq. 3.2.3), and hence the regression of  $X$  on  $\tau$  is linear. Also the regression function of  $E$  on  $\tau$ ,  $\mathcal{E}(E | \tau) = 0$ , is linear. The regression function of  $T$  on  $x$ , however, is generally not linear (see Sections 21.6 and 21.8). This lack of symmetry results from the particular relationship  $E$  has to  $T$  but not to  $X$ .

When questions arise that concern the “true” ability of individual examinees rather than differences in group means, regression estimates of true scores become quantities of interest. Then, regardless of the linearity or lack of linearity of the various regression functions of the classical test theory model, it is of interest to consider the various linear minimum mean squared error *regression functions*.\* For arbitrary random variables  $Y$  and  $Z$ , we define the linear minimum mean squared error *regression function* for the regression of  $Y$  on  $z$  as the

---

\* A review of regression and correlation theories is given in Sections 12.2 through 12.4.

linear function  $R(Y | z)$  which minimizes

$$\mathcal{E}\mathcal{E}[R^*(Y | z) - Y]^2,$$

where  $R^*(Y | z)$  is an arbitrary linear function of  $z$ , and expectation is taken over both  $Y$  and  $z$ . For simplicity we shall henceforth refer to the linear minimum mean squared error regression function as the *linear regression function*. The mean squared error is minimized when

$$R^*(Y | Z) = R(Y | Z),$$

where

$$R(Y | z) = \alpha + \beta z, \quad \alpha = \mu_Y - \beta \mu_Z \quad \text{and} \quad \beta = \rho_{YZ}(\sigma_Y / \sigma_Z).$$

The values  $\alpha$  and  $\beta$  are respectively the intercept and slope of the linear regression function, and  $\beta$  is called the *regression coefficient* of  $Y$  on  $z$ .

The regression coefficient  $\beta_{XT}$  of observed score on true score is then given by

$$\beta_{XT} = \rho_{XT} \frac{\sigma_X}{\sigma_T}.$$

But  $\rho_{XT} = \sigma_T / \sigma_X$ , hence

$$\beta_{XT} = 1, \tag{3.7.1}$$

and the linear regression function (in this case, the true regression function) is given by

$$R(X | \tau) = \tau, \tag{3.7.1a}$$

since

$$\alpha = \mu_X - \frac{\sigma_T}{\sigma_X} \frac{\sigma_X}{\sigma_T} \mu_T = 0.$$

Similarly the regression coefficient of true score on observed score is given by

$$\begin{aligned} \beta_{TX} &= \rho_{XT} \frac{\sigma_T}{\sigma_X} = \frac{\sigma_T^2}{\sigma_X^2} \\ &= \rho_{XX'}. \end{aligned} \tag{3.7.2}$$

The regression coefficient of true score on observed score is equal to the reliability of the test. The linear regression function is given by

$$R(T | x) = \rho_{XX'}x + (1 - \rho_{XX'})\mu_X. \tag{3.7.2a}$$

As Kelley (1947, p. 409) pointed out:

This is an interesting equation in that it expresses the estimate of true ability as a weighted sum of two separate estimates—one based upon the individual's observed score,  $X_1$ , and the other based upon the mean of the group to which he belongs, . . . If the test is highly reliable, much weight is given to the test score and little to the group mean, and vice versa.

The regression coefficient of one parallel measurement on a second is given by

$$\beta(X' | X) = \rho(X, X) \frac{\sigma(X')}{\sigma(X)} = \rho_{XX'}, \quad (3.7.3)$$

which is identical with the regression coefficient of true score on observed score. The linear regression function of one parallel measurement on another is given by

$$R(X' | x) = \rho_{XX'}x + (1 - \rho_{XX'})\mu_X, \quad (3.7.3a)$$

a result identical with (3.7.2a). The residual variances corresponding to the regression functions (3.7.2a) and (3.7.3a) are given as functions (3.8.4) and (3.8.5). The linear regression function of a true score on the true score of a distinct measurement when the two true-score variances are equal has the similar form

$$R(T_X | \tau_Y) = \rho(T_X, T_Y)\tau_Y + [1 - \rho(T_X, T_Y)]\mu(T_X). \quad (3.7.3b)$$

If  $X$  and  $Y$  are parallel, then  $\rho(T_X, T_Y) = 1$  and  $R(T_X | \tau_Y) = \tau_Y$ . Also, using (3.3.10), the regression coefficient of the error score on the observed score is given by

$$\beta_{EX} = \rho_{EX} \frac{\sigma_E}{\sigma_X} = 1 - \rho_{XX'}, \quad (3.7.4)$$

and the linear regression function of  $E$  on  $X$  is given by

$$R(E | x) = (1 - \rho_{XX'})(x - \mu_X). \quad (3.7.4a)$$

While we may seldom have need to predict an error score from an observed score, this formula does have some conceptual value. Note that if  $x > \mu_X$ , then the predicted value for  $E$  is positive. This corresponds with formula (3.7.2a) which gives a predicted value of  $T$  that is less than  $x$  when  $x > \mu_X$ . For example, suppose  $\mu_X = 51$  and  $\rho_{XX'} = 0.80$ , and we observe  $x_{ga} = 60$ . Then  $R(\tau_{ga} | 60) = (0.8)(60) + (0.2)(51) = 58.2$ . Also

$$R(e_{ga} | 60) = (0.20)(60 - 51) = 1.8.$$

### 3.8 Errors of Measurement, Estimation, and Prediction

Since the emphasis in the classical model is on the measurement of ability as indicated by the true score of a person on one or more measurements and the comparison of such estimates among examinees, the error random variable of primary interest is the variable  $E$  given in Section 3.1 and defined more precisely in Chapter 2. This error random variable is defined to be the difference between the observed-score random variable and the true-score random variable over persons  $a \in \mathcal{P}$ . It is the error involved when, for a randomly selected person, we take the person's observed score as a measure or estimate of his true score. In symbols we write

$$E = X - T, \quad (3.8.1)$$

and call the random variable  $E$ , taking values  $e$ , the *error of measurement*. If we take a single observation on a single person and take  $x$  to be an estimate of  $\tau$ , then  $e$  is a measure of the discrepancy between the estimate and the actual value of the true score. Since the estimate in this case is simply the measurement, we naturally refer to  $E$  as the error of measurement.

If, for the population  $\mathcal{P}$ , the quantities  $\rho_{XX'}$  and  $\mu_X$  are known and an observed score  $x_a$  is obtained for some person  $a$ , then the linear regression estimate of  $\tau_a$  is given by Eq. (3.7.2a), and this equation may be used to estimate the person's true score. If, in other words, the additional information  $\mu_X$  and  $\rho_{XX'}$  is known, a new estimate of  $\tau_a$  may be obtained from the linear regression function. This estimate is not the same as the estimate  $x_a$  unless the reliability of the measurement is unity.

The difference between the linear regression estimate and the true score itself may be written as

$$\epsilon = \rho_{XX'}(X - \mu_X) - (\tau - \mu_X) \quad (3.8.2)$$

and called the error of estimating true score from observed score or, more simply, the *error of estimation*.

If, for the population  $\mathcal{P}$ ,  $\rho_{XT}$  is known and an observed score  $x_{1a}$  is obtained for some person, then the linear regression estimate of a parallel measurement  $x_{2a}$  is given by (3.7.3a). The difference between the predicted value given by this equation and the actual value of the second parallel measurement may be written as

$$\Delta = [\rho_{XX'}X_1 + (1 - \rho_{XX'})\mu_X] - X_2 \quad (3.8.2a)$$

and called the *error of predicting an observed score* or, more simply, the *error of prediction*.

One measure of the magnitude of each of these errors is its standard deviation. It will prove of interest to examine the relative magnitude of these errors. In (3.3.9) above, the standard deviation of the errors of measurement, called the *standard error of measurement*, has already been given as

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}. \quad (3.8.3)$$

The standard deviation of the errors of estimation of true score or, more simply, the *standard error of estimation*, is given by the well-known formula

$$\sigma_\epsilon = \sigma_T \sqrt{1 - \rho_{XT}^2}. \quad (3.8.4)$$

We may verify this by noting that

$$\begin{aligned} \sigma_\epsilon &= \sigma[\tau - \mu_X - \rho_{XX'}X + \rho_{XX'}\mu_X] = \sigma[\tau - (\sigma_T^2/\sigma_X^2)X] \\ &= \sqrt{\sigma_T^2 - 2 \frac{\sigma_T^2}{\sigma_X^2} \sigma_{XT} + \frac{\sigma_T^2}{\sigma_X^2} \sigma_X^2} = \sqrt{\sigma_T^2(1 - 2\rho_{XT}^2 + \rho_{XT}^2)} = \sigma_T \sqrt{1 - \rho_{XT}^2}. \end{aligned}$$

If  $X$  and  $X'$  are taken as parallel measurements, then (3.8.4) may be written as

$$\sigma_\epsilon = \sigma_X \sqrt{\rho_{XX'}} \sqrt{1 - \rho_{XX'}}. \quad (3.8.4a)$$

The standard deviation of the errors of prediction or, more simply, the *standard error of prediction*, is given by

$$\sigma_\Delta = \sigma_X \sqrt{1 - \rho_{XX'}^2}. \quad (3.8.5)$$

The proof is left as an exercise for the reader. The reader may then easily verify that

$$\sigma_\Delta \geq \sigma_E \geq \sigma_\epsilon, \quad (3.8.6)$$

with equality holding if and only if  $\rho_{XX'} = 0$  or 1. It should be noted that each of the error standard deviations  $\sigma_E$ ,  $\sigma_\epsilon$ , and  $\sigma_\Delta$  decreases to zero as the reliability of the measurement increases to one. Thus the accuracy of measuring and estimating true scores or predicting parallel observed scores is an increasing function of reliability.

Suppose we have  $x_{ga} = 90$  and know that  $\sigma^2(X) = 25$ ,  $\rho_{XX'} = 0.80$ , and  $\mu_X = 100$ . We may then determine that

- 1)  $\sigma_T^2 = 20$  and  $\sigma_E^2 = 5$ ;
- 2) the regression estimate of true score or of a parallel measurement is

$$\rho_{XX'}x + (1 - \rho_{XX'})\mu_X = (0.80)(90) + (0.20)(100) = 92;$$

- 3) the standard errors of measurement, estimation, and prediction (as defined in Eqs. (3.8.3), (3.8.4a), and (3.8.5) are

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}} \doteq 5(0.447) = 2.24,$$

$$\sigma_\epsilon = \sigma_X \sqrt{1 - \rho_{XX'}} \sqrt{\rho_{XX'}} = 5(0.44)(0.89) = 1.958,$$

$$\sigma_\Delta = \sigma_X \sqrt{1 - \rho_{XX'}^2} = 5(0.6) = 3.$$

For expository purposes, we assume error variance homogeneity, and determine the corresponding approximate 90% Chebychev confidence intervals [ $k = \sqrt{10}$  in (3.4.1)] to be

$$|\tau_{ga} - 90| \leq 7.08, \quad |\tau_{ga} - 92| \leq 6.19, \quad |x_{g'a} - 92| \leq 9.48.$$

One may ask, heuristically, why the standard error of estimate is less than the standard error of measurement. The reason is that the regression estimate of true score uses some information that the measurement estimate does not, namely, the group mean and group reliability. Thus with linear regression theory we use information about the group, when it is available, to modify and improve our estimate of the true score of a person. A striking feature of this

theory is exemplified in Exercises 3.21 and 3.26, where we find that for  $\tau$ -equivalent measurements it is possible to have  $R(T_1 | x_1) > R(T_2 | x_2)$  when  $x_2 > x_1 > \mu_X$ , if  $\rho(X_1, X'_1) > \rho(X_2, X'_2)$ .

### 3.9 Attenuation Formulas

We now turn our attention to a problem which is of fundamental importance for theoretical psychologists and to which the classical test theory model is able to provide a very useful solution. In fact this problem is one which first motivated the development of test theory as a distinct discipline.

If a psychologist wishes to determine the relationship between two theoretical psychological variables, or *latent traits*, he may construct scales to measure them. If the relationship between these scales is linear, then the correlation coefficient indicates the measure of association between the scales. These scales contain error, however, and hence the correlation between the scales is less than the correlation between the traits. If, however, it can be presumed as a reasonable approximation that the true scores on the measurements can be taken to be the traits in question, then an attenuation formula can be used to compute the true correlation between the traits.

As a preliminary, consider two distinct measurements  $X$  and  $Y$  with true scores  $T_X$  and  $T_Y$ . The relation

$$\rho(X, T_Y) = \rho(X, T_X)\rho(T_X, T_Y) \quad (3.9.1)$$

is easily established by evaluation as follows. From (3.9.1) we have

$$\frac{\sigma(X, T_Y)}{\sigma(X)\sigma(T_Y)} = \frac{\sigma(X, T_X)}{\sigma(X)\sigma(T_X)} \frac{\sigma(T_X, T_Y)}{\sigma(T_X)\sigma(T_Y)}.$$

Since  $\sigma(X, T_X) = \sigma^2(T_X)$  and  $\sigma(X, T_Y) = \sigma(T_X, T_Y)$ , we have

$$\frac{\sigma(T_X, T_Y)}{\sigma(X)\sigma(T_Y)} = \frac{\sigma^2(T_X)}{\sigma(X)\sigma(T_X)} \frac{\sigma(T_X, T_Y)}{\sigma(T_X)\sigma(T_Y)},$$

and the identity is established. Clearly, then, since  $\sigma^2(T) > 0$ , we have

$$\rho(X, T_X) > 0, \quad (3.9.1a)$$

a statement whose proof we leave as an exercise (3.10) for the reader. Furthermore, since  $\rho(X, T_Y)$  and  $\rho(T_X, T_Y)$  must have the same sign, we have

$$\rho(X, T_X) \geq \rho(X, T_Y). \quad (3.9.1b)$$

*A measurement correlates at least as highly with its own true score as it does with the true score on a second measurement.*

Similarly we have

$$|\rho(X, T_Y)| \geq |\rho(X, Y)|. \quad (3.9.1c)$$

*Any measurement correlates with the true score on a second measurement at least as highly as it does with the observed score on the second measurement.* In fact we may write

$$|\rho(T_X, T_Y)| \geq |\rho(X, T_Y)| \geq |\rho_{XY}|. \quad (3.9.1d)$$

The remaining part of (3.9.1d) may be obtained by establishing the identity

$$\rho_{XY} = \rho(X, T_Y)\rho(Y, T_Y). \quad (3.9.2)$$

Similarly,

$$\rho_{XY} = \rho(Y, T_X)\rho(X, T_X). \quad (3.9.2a)$$

We now derive the important “attenuation formulas”. Substituting (3.9.1) into (3.9.2) we have

$$\rho_{XY} = \rho(X, T_X)\rho(T_X, T_Y)\rho(Y, T_Y). \quad (3.9.3)$$

Since, to avoid triviality, we assume generally that  $\sigma_E > 0$  and  $\sigma_T > 0$  and hence that  $\rho(X, T) > 0$ , we may rewrite (3.9.1) and (3.9.3) and obtain

$$\rho(T_X, T_Y) = \frac{\rho(X, T_Y)}{\rho(X, T_X)}, \quad (3.9.4)$$

$$\rho(T_X, T_Y) = \frac{\rho(X, Y)}{\rho(X, T_X)\rho(Y, T_Y)}. \quad (3.9.5)$$

Finally, given pairs of parallel measurements  $X, X'$  and  $Y, Y'$ , from (3.3.5b) and (3.9.5) we have

$$\rho(T_X, T_Y) = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}}. \quad (3.9.6)$$

*Equation (3.9.6) gives the correlation between true scores in terms of the correlation between observed scores and the reliability of each measurement.*

This formula may be interpreted as giving the correlation between the psychological constructs being studied in terms of the observed correlation of the measures of these constructs and the reliabilities of these measures. Also from (3.9.2) we have

$$\rho(X, T_Y) = \frac{\rho_{XY}}{\sqrt{\rho_{YY'}}}. \quad (3.9.7)$$

*Equation (3.9.7) gives the correlation between the observed score on one measurement and the true score on a second measurement in terms of the correlation between the observed scores on the two measurements and the reliability of the second measurement.*

Also the correlation between true scores is an upper bound on the correlation between observed scores, as stated in Eq. (3.9.5). For example, suppose the correlation between  $X$  and  $Y$  is 0.60, the reliability of  $X$  is 0.81, and the reli-

ability of  $Y$  is 0.64. Then the maximum correlation between  $X$  and  $Y$  that can be obtained by making the predictor  $X$  more reliable is

$$\rho(T_X, Y) = \frac{0.60}{\sqrt{0.81}} \doteq 0.67.$$

The maximum correlation obtainable by making the criterion  $Y$  more reliable is

$$\rho(X, T_Y) = \frac{0.60}{\sqrt{0.64}} \doteq 0.75.$$

The maximum correlation obtainable by making both the predictor and the criterion more reliable is

$$\rho(T_X, T_Y) = \frac{0.60}{\sqrt{(0.64)(0.81)}} \doteq 0.83.$$

In the next chapter we shall obtain these same formulas through a study of the effects that the lengthening of a test has on correlation.

Equations (3.9.6) and (3.9.7) are often referred to as *attenuation formulas*. The idea is that the correlation between observed scores is less than the correlation between corresponding true scores because the former correlation is *attenuated* by the unreliability of the measurements. If the reliabilities of the measurements are known, then these formulas may be used to compute the *disattenuated* correlations, i.e., the correlations between the corresponding true scores. Attenuation theory is one important justification for the emphasis that classical theory has placed on the concept of reliability.

The prediction of college grades is limited by a number of factors, including (i) the unreliability of the grades, which limits all test-criteria correlations, and (ii) changes in persons and situations over time, which cause the correlations between tests and grades to decrease over time. Attenuation theory can be used to argue effectively against the notion that, because of these limitations, tests have gone about as far as they can go in predicting college grades. Some illustrative data prepared by Turnbull (1966) for this purpose are summarized briefly in Table 3.9.1.

**Table 3.9.1**

Hypothetical intercorrelations indicating (i) typical fallible test scores and fallible course grades, and (ii) corresponding maximum correlations of a perfect predictor with fallible grades

	1.	2.	3.	4.	5.
1. Freshman test scores	(.90)	.87	.82	.77	.73
2. Freshman grades	.53	(.75)	.71	.67	.63
3. Sophomore grades	.50	.71	(.75)	.71	.67
4. Junior grades	.47	.67	.71	(.75)	.71
5. Senior grades	.44	.63	.67	.71	(.75)

The major diagonal elements (in parentheses) are reliabilities, the below-diagonal elements are typical intercorrelations. The above-diagonal elements represent the intercorrelations that would be found if the freshman test score were a perfect predictor of freshman grades except for the unreliability of the freshman grades. For example, the correlation between the perfect predictor and freshman grades is the attenuated correlation

$$\rho(T_X, Y) = \rho(T_X, T_Y)\sqrt{\rho_{YY'}} = (1.00)(\sqrt{0.75}) \doteq 0.87.$$

From the hypothetical "typical" data given in Table 3.9.1 and from the attenuation theory computations also given there, Turnbull has computed that the percentages of variance of course grades accounted for by the fallible test scores are 0.28, 0.25, 0.22, and 0.19, respectively, for the four college years, while the corresponding percentages for the "perfect predictor" would be 0.75, 0.67, 0.59, and 0.53. Thus, apparently, there is plenty of room for improvement in testing methods.

The following expression represents a further major justification for the study of reliability theory. From (3.9.1b) and (3.9.1c) we have

$$\rho_{XY} \leq \rho(X, T_X) = \sqrt{\rho_{XX'}}. \quad (3.9.8)$$

*The validity of a test with respect to any criterion cannot exceed the index of reliability.* This result has great significance in establishing the importance of reliability theory, for unless the index of reliability of a test is sufficiently high, validity cannot be high for *any* criterion. Of course, it does not follow that a high index of reliability guarantees or even suggests high validity; indeed it does not. High reliability is a necessary, but not a sufficient, condition for high validity. We shall discuss this point in more detail in Chapters 15, 16, and 20. It is also worth pointing out that the inequality of (3.9.8) also states that the index of reliability cannot be less than any validity coefficient of a test.

Now suppose that  $\rho_{XY} > \rho_{XX'}$ , that is, that the validity of  $X$  with respect to  $Y$  is greater than the reliability of  $X$  (though, of course, not greater than the index of reliability). Then  $\rho(Y, T_X) > \rho(X, T_X)$ . For if  $\rho_{XY} > \rho_{XX'} = \rho^2(X, T_X)$ , then  $\rho_{XY}/\rho(X, T_X) > \rho(X, T_X)$ . Thus, if  $\rho_{XY} > \rho_{XX'}$ , using (3.9.2a) we obtain

$$\rho(Y, T_X) > \rho(X, T_X). \quad (3.9.9)$$

*If the validity of one measurement with respect to a second is greater than the reliability of the first, then the second observed score correlates more highly with the first true score than does the first observed score.*

That the validity of a test (with respect to any other test) may exceed its reliability may at first appear contraintuitive. The paradox, however, results from the nonparallel character of the definitions of reliability and validity, the latter being defined as a coefficient of correlation, the former as a squared corre-

lation. Since  $|\rho| \leq 1$ , we must have  $\rho^2 \leq |\rho|$ . Cases in which validity exceeds reliability can occur when the predictor measure is quite reliable and the criterion measure is relatively unreliable. The reader should carefully note the difference between the statements accompanying (3.9.1b) and (3.9.9).

Attenuation theory is central to test theory and practice. However, there are substantial difficulties in applying this theory. These difficulties are discussed in Section 6.5. An important area of application of attenuation theory is that of partial correlation. While zero-order disattenuated correlations always have the same sign as the observed correlation, a change in sign is sometimes encountered when partial correlations are disattenuated (see Exercise 13.26).

Many psychological experiments are of the form, pretest, treatment, posttest. The most basic question in such experiments is the estimation of the true mean gain for the group following treatment. Since mean true score is equal to mean observed score, it follows that the difference between the mean observed posttest score and the mean observed pretest score is the regression estimate of the difference of the corresponding mean true scores.

It is sometimes of interest to estimate an individual examinee's true gain from his observed gain. This is a more complicated problem and we shall treat it in detail in the next section, using an extension of the regression theory developed in Section 3.7. A third question, which is occasionally of interest, is the determination of the correlation of the true gain with the initial true score. The solution of this problem is based on attenuation theory and therefore is treated in this section.

Let  $X_1$ , with true score  $T_1$ , and  $X_2$ , with true score  $T_2$ , be the measurements before and after treatment, respectively. Then  $X_2 - X_1$  is the observed-difference measurement and  $T_2 - T_1$  is the true-difference measurement. These differences are often referred to as change or gain measurements. Our problem is to express the correlation  $\rho(T_2 - T_1, T_1)$  in terms of observables. We must exercise some care in the derivation since  $X_2 - X_1$  and  $X_1$  are *not* experimentally independent. The error  $E_1$  of the measurement  $X_1$  is necessarily a part of the difference measurement  $X_2 - X_1$  and necessarily occurs with opposite signs in the two measurements. Thus there tends to be a spurious negative correlation between pretest and observed gain. This fact was first pointed out by Sir Godfrey Thomson (1924). Further related problems in the measurement of growth have been discussed by Lord (1958, 1963).

The required result is given in

**Theorem 3.9.1.** The true correlation of gain with initial score is given by

$$\begin{aligned} & \rho(T_2 - T_1, T_1) \\ &= \frac{\sigma(X_2) \left[ \rho(X_1, X_2) - \frac{\sigma(X_1)}{\sigma(X_2)} \rho(X_1, X'_1) \right]}{[\rho(X_1, X'_1)]^{1/2} [\rho(X_2, X'_2) \sigma^2(X_2) - 2\rho(X_1, X_2) \sigma(X_1) \sigma(X_2) + \rho(X_1, X'_1) \sigma^2(X_2)]^{1/2}}. \end{aligned}$$

*Proof*

$$\begin{aligned} \rho(T_2 - T_1, T_1) &= \frac{\sigma(T_1, T_2) - \sigma^2(T_1)}{\sigma(T_2 - T_1)\sigma(T_1)} = \frac{\sigma(T_2)\rho(T_1, T_2) - \sigma(T_1)}{[\sigma^2(T_2) + \sigma^2(T_1) - 2\sigma(T_1, T_2)]^{1/2}} \\ &= \frac{\sigma(X_2)\sqrt{\rho(X_2, X'_2)} \frac{\rho(X_1, X_2)}{\sqrt{\rho(X_1, X'_1)\rho(X_2, X'_2)}} - \sigma(X_1)\sqrt{\rho(X_1, X'_1)}}{[\sigma^2(X_2)\rho(X_2, X'_2) + \sigma^2(X_1)\rho(X_1, X'_1) - 2\sigma(X_1, X_2)]^{1/2}}, \end{aligned}$$

from which the result follows immediately.  $\square$

The reader should note that the attenuation formula (3.9.6) could *not* be used directly to express  $\rho(T_2 - T_1, T_1)$  in terms of observables but could be used to express  $\rho(T_2, T_1)$  in terms of observables.

### 3.10 Elementary Models for Inferring True Change

The most general regression model for inferring true change from observed change makes no assumptions about the relative magnitudes of the observed-score variances, error variances, or reliabilities of  $X_1$  and  $X_2$ . Thus, if we wish to infer the true gain  $G = T_2 - T_1$  from the observed gain  $X_2 - X_1$  for any specified individual, we must use the full regression model. The general form is

$$R(G | x_1, x_2) = (\mu_2 - \mu_1) + \beta_1(x_1 - \mu_1) + \beta_2(x_2 - \mu_2). \quad (3.10.1)$$

A simple method for obtaining the constants was sketched by McNemar (1958). The regression coefficients  $\beta_1$  and  $\beta_2$  are

$$\beta_1 \equiv \beta_{G1 \cdot 2} = \frac{\sigma_{G \cdot 12}}{\sigma_{1 \cdot G2}} \rho_{G1 \cdot 2} \quad (3.10.2)$$

and

$$\beta_2 \equiv \beta_{G2 \cdot 1} = \frac{\sigma_{G \cdot 12}}{\sigma_{2 \cdot G1}} \rho_{G2 \cdot 1}. \quad (3.10.3)$$

The squared partial correlations are

$$\rho_{G1 \cdot 2}^2 = \frac{(\rho_{G1} - \rho_{12}\rho_{G2})^2}{(1 - \rho_{12}^2)(1 - \rho_{G2}^2)} \quad (3.10.4)$$

and

$$\rho_{G2 \cdot 1}^2 = \frac{(\rho_{G2} - \rho_{12}\rho_{10})^2}{(1 - \rho_{12}^2)(1 - \rho_{G1}^2)}, \quad (3.10.5)$$

where

$$\rho_{G1} \equiv \rho(G, X_1) = \frac{\sigma(X_1)\rho(X_1, X'_1) - \sigma(X_2)\rho(X_2, X'_1)}{\sigma(G)}, \quad (3.10.6)$$

$$\rho_{G2} \equiv \rho(G, X_2) = \frac{\sigma(X_2)\rho(X_2, X'_2) - \sigma(X_1)\rho(X_1, X'_2)}{\sigma(G)}. \quad (3.10.7)$$

The partial variances are

$$\sigma_{G \cdot 12}^2 = \sigma_G^2(1 - \rho_{G \cdot 12}^2), \quad (3.10.8)$$

$$\sigma_{1 \cdot G2}^2 = \sigma_1^2(1 - \rho_{1 \cdot G2}^2), \quad (3.10.9)$$

$$\sigma_{2 \cdot G1}^2 = \sigma_2^2(1 - \rho_{2 \cdot G1}^2). \quad (3.10.10)$$

The variance  $\sigma_{G \cdot 12}^2$  is the residual variance when  $G$  is predicted linearly from  $X_1$  and  $X_2$ . The variance of  $G$  is

$$\sigma_G^2 = \sigma^2(X_2)\rho(X_2, X'_2) + \sigma^2(X_1)\rho X_1, X'_1 - 2\rho(X_2, X_1)\sigma(X_2)\sigma(X_1). \quad (3.10.11)$$

The multiple correlations are

$$\rho_{G \cdot 12}^2 = \frac{\rho_{G1}^2 + \rho_{G2}^2 - 2\rho_{G1}\rho_{G2}\rho_{12}}{1 - \rho_{12}^2}, \quad (3.10.12)$$

$$\rho_{1 \cdot G2}^2 = \frac{\rho_{G1}^2 + \rho_{12}^2 - 2\rho_{G1}\rho_{12}\rho_{G2}}{1 - \rho_{G2}^2}, \quad (3.10.13)$$

$$\rho_{2 \cdot G1}^2 = \frac{\rho_{G2}^2 + \rho_{12}^2 - 2\rho_{G2}\rho_{12}\rho_{G1}}{1 - \rho_{G1}^2}. \quad (3.10.14)$$

The above formulas can be obtained from the general formulas for partial correlations and variances and for multiple correlations given in Chapter 12.

If it is reasonable to assume that the (average) error variance for the initial and final scores are equal, then we may use the much simpler regression estimate derived by Lord (1956). The regression equation is

$$R(G | x_1, x_2) = (\mu_2 - \mu_1) + \beta_{G2 \cdot 1}^*(x_2 - \mu_2) + \beta_{G1 \cdot 2}^*(x_1 - \mu_1), \quad (3.10.15)$$

where

$$\beta_{G2 \cdot 1}^* = \frac{\rho_{22'} - \rho_{12}^2 - \{(1 - \rho_{11'})\sigma_1\rho_{12}\}/\sigma_2}{1 - \rho_{12}^2} \quad (3.10.16)$$

and

$$\beta_{G1 \cdot 2}^* = \frac{\{(1 - \rho_{22'})\rho_{12}\sigma_2\}/\sigma_1 - \rho_{11'} + \rho_{12}^2}{1 - \rho_{12}^2}, \quad (3.10.17)$$

$\rho_{11'}$  and  $\rho_{22'}$  being reliabilities. The residual variance is given by (3.10.8).

Suppose  $\mu_1 = 100$ ,  $\mu_2 = 110$ ,  $x_1 = 115$ ,  $x_2 = 112$ ,  $\rho(X_1, X'_1) = 0.8$ ,  $\rho(X_1, X_2) = 0.7$ , and  $\sigma^2(X_1) = \sigma^2(X_2) = 10$ . One can readily determine that  $R(G | x_1, x_2) = 5\frac{2}{3}$ . Thus we have a situation in which the observed change is negative but the inferred true change is positive. The “explanation” of this paradox is that the first observed score is estimated to have a large positive error component. Since  $X_1 > \mu_1$ , we expect the second observed score to be lower even if there is no true change.

McNemar has suggested that the regression function  $R(G | x_2 - x_1)$  be used as an approximation to the regression function  $R(G | x_2, x_1)$ . We leave the derivation of this regression function as an exercise for the reader. The reader should note, however, that the residual variance in this case is generally greater than that given by (3.10.8), in which  $G$  is predicted from the best linear combination of  $x_1$  and  $x_2$ . For most careful work, the general formulas given at the beginning of this section should be used.

Since gain scores are notoriously unreliable, one should never estimate the true gains of individuals without at the same time estimating the variance of these estimates (3.10.8).

It is important to note that the method of this section is appropriate for estimating the true gain for individuals, but *not* for estimating group parameters. If one wants to estimate the correlation of true gain with some outside variable  $Z$ , for example, one should *not* estimate the true gain for each individual and then correlate these estimates with  $Z$ . Instead one should correlate observed change  $X_2 - X_1$  with  $Z$  and correct this correlation for the attenuation due to the unreliability of  $X_2 - X_1$ . The reliability coefficient of  $X_2 - X_1$  is given by

$$\frac{\rho(X_2, X'_2)\sigma^2(X_2) + \rho(X_1, X'_1)\sigma^2(X_1) - 2\rho(X_1, X_2)\sigma(X_1)\sigma(X_2)}{\sigma^2(X_2 - X_1)}. \quad (3.10.18)$$

The reader can easily verify this.

### Exercises

- 3.1. The reader who has had little prior exposure to test theory may feel himself confronted by a massively long string of formulas, with perhaps all too little indication of the relative importance of the various formulas. Undoubtedly he will want to commit some formulas to memory—but which ones? The choice of formulas to be memorized should be based on the ease with which they can be used to obtain any others that are needed. For the material of this chapter, we recommend as a very bare minimum that the reader memorize the following basic assumptions, definitions, and results:

$$\begin{aligned} X &= T + E, \quad \mathcal{E}(E | \tau) = 0, \quad \rho(E_g, E_h) = 0, \\ \sigma_X^2 &= \sigma_T^2 + \sigma_E^2, \quad \rho_{XT}^2 = \rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}, \\ \sigma(X_1, X_2) &= \sigma(X_1, T_2) = \sigma(T_1, T_2) = \sigma(T_1, X_2). \end{aligned}$$

Without referring to the text, or to any notes, the reader should see just how much of the material of this chapter he can reproduce from these formulas. He will find that almost all other formulas can be obtained by one-, two-, or three-step derivations.

- 3.2. Let  $X_1, X_2, X_3, \dots$  be parallel tests. Using only equations preceding (3.3.6a), show that  $\rho(X_1, X_2) = \rho(X_1, X_3) = \rho(X_2, X_3) = \dots$ .
- 3.3. Let  $X_1, X_2, X_3, \dots$  be parallel tests, and  $Z$  a distinct test. Using only equations preceding (3.3.6b), show that  $\rho(X_1, Z) = \rho(X_2, Z) = \rho(X_3, Z) = \dots$ .
- 3.4. Show by direct evaluation that  $\sigma(T_1, X_2) = \sigma(X_1, T_2) = \sigma(T_1, T_2) = \sigma(X_1, X_2)$ .
- 3.5. Show that  $\sigma(X_1, E_1) = \sigma^2(E_1)$ .
- 3.6. Show that results (3.5.5) and (3.5.6) hold for  $\tau$ -equivalent measurements.
- 3.7. For measurements  $X_1, X_2$ , and  $X_3$  which are  $\tau$ -equivalent but not parallel, show that

$$\sigma(X_1, X_2) = \sigma(X_1, X_3) = \sigma(X_2, X_3),$$

but that

$$\rho(X_1, X_2) \neq \rho(X_1, X_3) \neq \rho(X_2, X_3).$$

- 3.8. Prove that  $\rho_{XE}^2 = \sigma_E^2/\sigma_X^2$ .

- 3.9. Show that the following linear regression coefficients and functions are as given:

$$\beta_{XE} = 1, \quad R(X | e) = \mu + e, \quad R(T | e) = \mu_T, \quad R(E | \tau) = 0.$$

- 3.10. Prove that  $\rho(X, T_X) > 0$ .

- 3.11. Show that

$$\rho_{XY} \leq \rho(X, T_X)\rho(Y, T_Y), \quad \rho_{XY} \leq \rho(X, T_X)\rho(T_X, T_Y),$$

and

$$\rho_{XY} \leq \rho(T_X, T_Y)\rho(Y, T_Y).$$

- 3.12. The marginal values in the following table are values of observed-score variances and reliability coefficients. Complete the body of the table by entering the corresponding true-score variances and error variances above and below the diagonals in the appropriate cells. The results  $\sigma_T^2 = \frac{1}{8}$ ,  $\sigma_E^2 = \frac{3}{8}$  have been entered for  $\sigma_X^2 = \frac{1}{2}$ ,  $\rho_{XX'} = \frac{1}{4}$ .

$\sigma_X^2$	$\rho_{XX'}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$
$\frac{1}{4}$						
$\frac{1}{2}$	$\frac{1}{8}$	$\frac{3}{8}$				
$\frac{2}{3}$						
1						
2						

- 3.13. Let  $Y_1, Y_2, \dots, Y_i, \dots$  and  $Z_1, Z_2, \dots, Z_j, \dots, Z_n$  be distinct measurements. Show that it is possible to construct distinct measurements

$$X_1, X_2, \dots, X_k, \dots, X_p$$

as linear functions of a finite number of the measurements  $Y_i$  such that, for all  $k', k'',$  and  $j$ :

- a)  $\mathcal{E}(X_{k'}) = \mathcal{E}(X_{k''}) = \mathcal{E}(X_k),$
- b)  $\sigma^2(X_{k'}) = \sigma^2(X_{k''}) = \sigma^2(X_k),$
- c)  $\rho(X_k, X_{k'}) = \rho(X_k, X_{k''}) = \rho(X_{k'}, X_{k''}),$
- d)  $\rho(X_{k'}, Z_j) = \rho(X_{k''}, Z_j) = \rho(X_k, Z_j), j = 1, 2, \dots, n.$

Also show that

- e)  $X_k$  and  $X_{k'}$  are *not* parallel according to the definition adopted in this book.

- 3.14. If a measurement procedure has reliability 0.90 and the mean true score is 100 and the observed score of a randomly selected person is 120, what is the regression estimate of his true score? (Answer: 118.) Suppose a second person has an observed score of 90. What is his estimated true score? What are the corresponding estimates for these two people if the reliability of the measurement procedure is 0.10?

- 3.15. If measurements  $X$  and  $Y$  have a correlation of 0.50 and each has a reliability of 0.50, what is the correlation between the true scores of  $X$  and  $Y$ ?

- 3.16. Suppose we observe  $x_{ga} = 100$ , and know that

$$\sigma(E_{ga}) = 10 \quad \text{and} \quad \sigma(X_{ga}) = 25.$$

What assumption would be necessary to use (3.4.1) to place a Chebychev confidence bound on  $\tau_{ga}$ ? Make this assumption; take  $k$  to be specified as 2 and write this interval explicitly. If person  $a$  is considered to have been sampled randomly, how will the Chebychev bound on  $\mu_X$  differ from that on  $\tau_{ga}$ ?

- 3.17. Derive the linear minimum mean squared error regression functions of  $T_1$  on  $x_2$ , and the corresponding residual variance. Compare this with the regression function of  $X_1$  on  $x_2$  and the corresponding residual variance.

- 3.18. If  $\rho_{01 \cdot 2}$  is the partial correlation between observed random variables  $X_0$  and  $X_1$  when the effect of  $X_2$  is partialled out, and  $\tilde{\rho}_{01 \cdot 2}$  is the partial correlation between  $X_0$  and  $X_1$  when the effect of  $T_2$  (the true score of  $X_2$ ) is partialled out, show that

$$\tilde{\rho}_{01 \cdot 2} = \frac{(\rho_{01} - \tilde{\rho}_{20}\tilde{\rho}_{12})^2}{(1 - \tilde{\rho}_{02}^2)(1 - \tilde{\rho}_{12}^2)} = \frac{[\rho_{01} - (\rho_{20}\rho_{12}/\rho_{22})]^2}{[1 - (\rho_{02}^2/\tilde{\rho}_{22})][1 - (\rho_{12}^2/\tilde{\rho}_{22})]}.$$

- 3.19. If the validities of the observed-score random variables  $X_1$  and  $X_2$  are 0.6 and 0.65, respectively, their intercorrelation is 0.7 and the reliability of  $X_2$  is 0.8. Compare  $\rho_{01 \cdot 2}^2$  and  $\tilde{\rho}_{01 \cdot 2}^2$ . (Answer:  $\rho_{01 \cdot 2}^2 = 0.072$ ,  $\tilde{\rho}_{01 \cdot 2}^2 = 0.005$ .)

- 3.20. Now suppose the intercorrelation between  $X_1$  and  $X_2$  is only 0.2. Compare  $\rho_{01 \cdot 2}^2$  and  $\tilde{\rho}_{01 \cdot 2}^2$ .

- 3.21. Suppose  $\mu_X = 100$ . For a given  $X$ , consider the observed scores and reliabilities in the margins of the following table. Use Eq. (3.7.2a) to complete the table and give the regression estimate of  $T$ .

## Reliability

	0.50	0.60	0.70	0.80	0.85	0.90	0.95
Observed score	85					73	
70							
80							
90							
95							
100							
105							
110			107				
120							119
130	115						

- 3.22. Verify Turnbull's findings as reported in Table 3.9.1.
- 3.23. Suppose  $X$  is a measure of one psychological trait and  $Y$  a measure of another, and that each measurement has a reliability 0.75. In a very large sample we observe a correlation  $\rho(X, Y) = 0.70$ , indicating that  $X$  accounts for less than 50% of the variance of  $Y$ . Using attenuation theory, what can we say about the two psychological traits being studied?
- 3.24. Let  $X_1$  and  $X_2$  be essentially  $\tau$ -equivalent measurements with true scores  $a_1 + T_1$  and  $a_2 + T_2$ . Find the linear minimum mean squared error regression function of  $T$  on  $x_1$  and  $x_2$ .
- 3.25. Suppose we have a measurement  $X$  and two populations  $\mathcal{P}_1$  and  $\mathcal{P}_2$ . Suppose that  $\mathcal{E}(X) = 100$  in  $\mathcal{P}_1$  and  $\mathcal{E}(X) = 110$  in  $\mathcal{P}_2$ , and also that an examinee in  $\mathcal{P}_1$  has an observed score of 106 and an examinee in  $\mathcal{P}_2$  has an observed score of 105. Assume that in both populations the reliability of the instrument is 0.80. Compute the regression estimate of each examinee's true score.
- 3.26. Assume  $\tau$ -equivalent measurements  $X_1$  and  $X_2$  with reliabilities  $\rho(X_1, X'_1) = 0.90$  and  $\rho(X_2, X'_2) = 0.80$ , and assume  $\mathcal{E}(X_1) = \mathcal{E}(X_2) = 100$ . Suppose one examinee receives a score of 120 on  $X_2$  and a second examinee receives a score of 119 on  $X_1$ . Which examinee do you estimate has the higher true score?
- 3.27. Show that  $\sigma_\Delta = \sigma_X \sqrt{1 - \rho_{XX'}^2}$  and hence that  $\sigma_\Delta \geq \sigma_E \geq \sigma_\epsilon$ . Give a heuristic explanation of why  $\sigma_\Delta \geq \sigma_\epsilon$ .
- 3.28. Derive the regression function (3.10.12).
- 3.29. Let  $X_1$  and  $X_2$  be measurements with true scores  $T_1$  and  $T_2$ , mean observed scores  $\mu_1$  and  $\mu_2$ , observed-score variances  $\sigma^2(X_1)$  and  $\sigma^2(X_2)$ , and reliabilities  $\rho(X_1, X'_1)$  and  $\rho(X_2, X'_2)$ . Derive the regression function of  $T_2 - T_1$  on  $x_2 - x_1$ .
- 3.30. Show that many of the results of this chapter based on parallel measurements can be obtained assuming only pseudoparallelism, provided the results are interpreted to apply only for sampling from the entire population.

- 3.31. Show that  $R(G | x_2 - x_1) = R(G | x_1, x_2)$  if

$$\sigma^2(X_1) = \sigma^2(X_2) = \sigma^2(X) \quad \text{and} \quad \rho(X_1, X'_1) = \rho(X_2, X'_2) = \rho(X, X').$$

- 3.32. Show that if the conditions of Exercise 3.31 hold, then

a)  $\rho^2(G, X_2 - X_1) = \frac{\rho(X, X') - \rho(X_1, X_2)}{1 - \rho(X_1, X_2)},$

b)  $\rho(X, X') > \rho(X_1, X_2),$

c)  $\rho^2(G \cdot X_1, X_2) = \rho^2(G \cdot X_2 - X_1) = \beta_{G2 \cdot 1}.$

- 3.33. a) Show that if the conditions of Exercise 3.31 hold, then  $R(G | x_1, x_2)$  and  $x_2 - x_1$  will have opposite signs if and only if

$$\rho(X_1, X_2) > \rho(X, X') + [1 - \rho(X, X')] \left( \frac{\mu_2 - \mu_1}{x_2 - x_1} \right).$$

- b) Show that the necessary and sufficient condition in (a) is equivalent to

$$\frac{\mu_2 - \mu_1}{x_2 - x_1} < \frac{\rho(X_1, X_2) - \rho(X, X')}{1 - \rho(X, X')}.$$

- c) Will the ratio  $(\mu_2 - \mu_1)/(x_2 - x_1)$  be positive or negative when  $x_2 - x_1$  and  $R(G | x_1, x_2)$  have opposite signs?

- 3.34. a) Suppose that the conditions of Exercise 3.31 hold, and that  $\mu_2 = \mu_1$ . Use Exercise 3.33 to show that  $R(G | x_1, x_2)$  and  $x_2 - x_1$  will have opposite signs if and only if  $\rho(X_1, X_2) > \rho(X, X')$ .  
 b) Is this condition possible under the given assumptions?

- 3.35. Assume  $\mathcal{E}X_1 = 100$ ,  $\mathcal{E}X_2 = 100$ ,  $\sigma^2(X_1) = \sigma^2(X_2) = 100$ ,  $\rho(X_1, X'_1) = \rho(X_2, X'_2) = 0.8$ , and  $\rho(X_1, X_2) = 0.7$ . Find the  $R(G | x_1, x_2)$  and residual variance if  $X_1 = 140$  and  $X_2 = 138$ .

- 3.36. Suppose we have two measurements,  $X_1$  and  $X_2$ , where  $X_1$  is a measurement before treatment and  $X_2$  is a measurement after treatment. Suppose further that  $\mathcal{E}(X_1) = 90$ ,  $\mathcal{E}(X_2) = 100$ ,  $\rho(X_1, X'_1) = 0.60$ ,  $\rho(X_2, X'_2) = 0.90$ , and  $\rho(X_1, X_2) = 0.55$  in population  $\mathcal{P}$ . Suppose finally that an examinee in  $\mathcal{P}$  has an observed score of 120 before treatment and 115 after treatment. Find

$$R(G | x_1, x_2).$$

- 3.37. Derive Eq. (3.8.5).

- 3.38. Derive the regression function  $R(G | x_1, x_2)$  and the associated residual variance.

## References and Selected Readings

GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.

KELLEY, T. L., *Fundamentals of statistics*. Cambridge: Harvard University Press, 1947.

KOUTSOPoulos, C. J., The mathematical foundations of classical test theory: an axiomatic approach, I. *Research Bulletin 62-17*. Princeton, N.J.: Educational Testing Service, 1962.

- KOUTSOPoulos, C. J., The mathematical foundations of classical test theory: an axiomatic approach, II. *Research Memorandum 64-3*. Princeton, N.J.: Educational Testing Service, 1964.
- LORD, F. M., The measurement of growth. *Educational and Psychological Measurement*, 1956, **16**, 421-437. See also *Errata*, ibid; 1957, **17**, 452.
- LORD, F. M., Further problems in the measurement of growth. *Educational and Psychological Measurement*, 1958, **18**, 437-454.
- LORD, F. M., Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: The University of Wisconsin Press, 1963, 21-38.
- McNEMAR, Q., On growth measurement. *Educational and Psychological Measurement*, 1958, **18**, 47-56.
- NOVICK, M. R., The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, **3**, 1-18.
- THOMSON, G., A formula to correct for the effect of errors of measurement on the correlation of initial values with gain. *Journal of Experimental Psychology*, 1924, **7**, 321-324.
- TURNBULL, W. W., Sources of error in the prediction of college grades. Unpublished manuscript. Princeton, N.J.: Educational Testing Service, 1966.

## CHAPTER 4

# COMPOSITE TESTS

### 4.1 Introduction

Most psychological tests are composed of items and often these items are grouped together into subtests. The total score that a person receives on a test is then usually computed by adding or averaging (in a weighted or unweighted manner) the scores on the individual items or subtests. In this chapter, and in Chapters 5, 9, 11, 13 through 20, and 23, we shall be concerned with models for tests composed of such additive parts.

The theory of measurement that we have presented can easily be extended to cover tests of this kind. We need only consider that each item, subtest, or test generates a measurement, that the subtest measurement is determined in an additive manner from the item measurements, and that the total test measurement is similarly determined from the subtest measurements. We refer to the total measurement as the *composite* measurement and its parts as the *component* measurements. Thus we shall be concerned with determining the characteristics of a composite test from the characteristics of its components. A major task of test theory is to determine what kinds and numbers of components must be put together to construct a composite having specified properties.

In a number of sections we shall be concerned with the *length of a test*. By the *length of a test* we shall mean the *number of components* comprising the composite test. In practice, the components are sometimes taken to be groups of items. In the simple case in which the items themselves are taken as the components, the length of the test is simply the number of items in the test. However, in another context, the length of a test (or measurement) may have a somewhat different meaning. For example, suppose we wish to measure a student's reading speed. In this case, the length of a test could be the time that the student is permitted to read, and his test score could be the number of words correctly read. Alternatively, the length of the test could be the number of words he is permitted to read, and his test score then could be the time taken to read these words. Models employing these differing interpretations of test length are given in Chapter 21.

In this chapter, we are concerned entirely with tests composed of discrete units, such as items. In Chapter 5, we shall present a more general model which covers both this case and that of a test whose length is determined by a contin-

uous time parameter. The models discussed in Chapter 21 are parametric models having continuous time parameters. In the next two sections of this chapter, we shall be concerned primarily with cataloguing some mathematical results which will be useful in succeeding work. Sections 4.3 and 4.4 deal with coefficient  $\alpha$ , a quantity of central importance to classical theory. Sections 4.2 through 4.5 deal with the case of unweighted composites. The more general case of weighted composites is covered in Sections 4.6 and 4.7.

## 4.2 Composite Measurements with Two Components

Let  $(Y_1, T_1, E_1)$  and  $(Y_2, T_2, E_2)$  be measurements taking values  $(y_1, \tau_1, e_1)$  and  $(y_2, \tau_2, e_2)$ . Let  $(X, T, E)$  be a composite measurement taking values  $(x, \tau, e)$  and defined by

$$X = Y_1 + Y_2. \quad (4.2.1)$$

Then it may easily be shown that, correspondingly,

$$T = T_1 + T_2, \quad E = E_1 + E_2 \quad (4.2.2)$$

must hold if  $T$  and  $E$  are to be the true and error scores corresponding to the observed score  $X$ . Then, denoting  $\mathcal{E}(Y_1) = \mu_1$ ,  $\mathcal{E}(Y_2) = \mu_2$ , and  $\mathcal{E}(X) = \mu$ , we have

$$\mathcal{E}(X) = \mathcal{E}(Y_1 + Y_2) = \mu_1 + \mu_2 = \mu. \quad (4.2.3)$$

The observed-, true-, and error-score variances for the composite tests are easily obtained. The variance of the composite observed score is given by

$$\begin{aligned} \sigma_X^2 &= \sigma^2(Y_1 + Y_2) = \sigma^2(Y_1) + \sigma^2(Y_2) + 2\sigma(Y_1, Y_2) \\ &= \sigma^2(Y_1) + \sigma^2(Y_2) + 2\rho(Y_1, Y_2)\sigma(Y_1)\sigma(Y_2). \end{aligned} \quad (4.2.4)$$

The variance of the composite true score is given by

$$\begin{aligned} \sigma_T^2 &= \sigma^2(T_1 + T_2) = \sigma^2(T_1) + \sigma^2(T_2) + 2\sigma(T_1, T_2) \\ &= \sigma^2(T_1) + \sigma^2(T_2) + 2\rho(T_1, T_2)\sigma(T_1)\sigma(T_2). \end{aligned} \quad (4.2.5)$$

The variance of the composite error score has the simpler form

$$\begin{aligned} \sigma_E^2 &= \sigma^2(E_1 + E_2) = \sigma^2(E_1) + \sigma^2(E_2) + 2\rho(E_1, E_2)\sigma(E_1)\sigma(E_2) \\ &= \sigma^2(E_1) + \sigma^2(E_2), \end{aligned} \quad (4.2.6)$$

since the correlation of error scores on distinct measurements is zero.

If  $Y$  and  $Y'$  are parallel measurements, the expressions for the composite observed, true, and error scores have the simpler forms

$$\sigma_X^2 = 2\sigma_Y^2[1 + \rho_{YY'}], \quad (4.2.7)$$

$$\sigma_T^2 = 4\sigma_T^2(T_1), \quad (4.2.8)$$

$$\sigma_E^2 = 2\sigma_E^2(E_1). \quad (4.2.9)$$

The proof of (4.2.9) is obvious, since  $\sigma^2(E_1) = \sigma^2(E_2)$ . The proof of (4.2.8) follows from the fact that if  $Y_1$  and  $Y_2$  are parallel measurements, then  $T_1 \equiv T_2$ , and hence  $\sigma(T_1, T_2) = \sigma^2(T_1)$ . The proof of (4.2.7) follows from the relation  $\sigma^2(Y_1) = \sigma^2(Y_2)$ .

Next we obtain the reliability of a composite measurement composed of two parallel measurements. Let  $Y_1, Y_2, Y_3$ , and  $Y_4$  be four parallel measurements, and let  $X = Y_1 + Y_2$  and  $X' = Y_3 + Y_4$ . The reader can easily verify that  $X$  and  $X'$  are parallel measurements. Then the reliability of  $X$  is given by  $\rho(Y_1 + Y_2, Y_3 + Y_4)$ . Using (3.5.1), we have

$$\begin{aligned} & \rho(Y_1 + Y_2, Y_3 + Y_4) \\ &= \frac{\sigma(Y_1 + Y_2, Y_3 + Y_4)}{\sigma(Y_1 + Y_2)\sigma(Y_3 + Y_4)} \\ &= \frac{\sigma(Y_1, Y_3) + \sigma(Y_2, Y_3) + \sigma(Y_1, Y_4) + \sigma(Y_2, Y_4)}{\{[\sigma^2(Y_1) + \sigma^2(Y_2) + 2\sigma(Y_1, Y_2)][\sigma^2(Y_3) + \sigma^2(Y_4) + 2\sigma(Y_3, Y_4)]\}^{1/2}}. \end{aligned}$$

However, since the observed-score variances and intercorrelations of all parallel tests are identical, we have

$$\rho(Y_1 + Y_2, Y_3 + Y_4) = \frac{4\sigma^2(Y_1)\rho(Y_1, Y_2)}{2\sigma^2(Y_1) + 2\rho(Y_1, Y_2)\sigma^2(Y_1)} = \frac{2\rho(Y_1, Y_2)}{1 + \rho(Y_1, Y_2)},$$

or

$$\rho_{XX'} = \frac{2\rho_{YY'}}{1 + \rho_{YY'}}, \quad (4.2.10)$$

where  $\rho_{YY'} \equiv \rho(Y_1, Y_2)$ . Equation (4.2.10) gives the reliability  $\rho_{XX'}$  of a composite test composed of two parallel component parts each having reliability  $\rho_{YY'}$ . This result is known as the Spearman-Brown formula for the reliability of a test of double length.

If  $0 < \rho_{YY'} < 1$ , then  $2/(1 + \rho_{YY'}) > 1$ , and hence

$$\frac{2\rho_{YY'}}{1 + \rho_{YY'}} = \rho_{XX'} > \rho_{YY'}.$$

The reliability of this composite test is greater than that of either component. If  $\rho_{YY'} = 1$ , then  $\rho_{XX'} = 1$ . Some values for the reliability of a test of double length as a function of reliability at unit length are given in Table 4.2.1.

**Table 4.2.1**

Reliability at double length as a function of reliability at unit length

Original reliability	.050	.100	.200	.300	.400	.500	.600	.700	.800	.900	.950
Reliability of test of double length	.095	.182	.333	.462	.571	.667	.750	.824	.889	.947	.974

### 4.3 Composite Measurements with $n$ Components

As a natural extension to the previous section, consider measurements  $(Y_i, T_i, E_i)$  taking values  $(y_i, \tau_i, e_i)$  for  $i = 1, 2, \dots, n$ . Let  $(X, T, E)$  be the composite measurement taking values  $(x, \tau, e)$  and defined by

$$X = \sum_{i=1}^n Y_i. \quad (4.3.1)$$

Then

$$E = \sum_{i=1}^n E_i, \quad (4.3.2)$$

$$T = \sum_{i=1}^n T_i. \quad (4.3.3)$$

Also

$$\mathcal{E}(X) = \mathcal{E}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \mu_i = \mu. \quad (4.3.4)$$

The expectations of  $T$  and  $E$  are given by

$$\mathcal{E}T = \sum_{i=1}^n \mathcal{E}T_i = \sum_{i=1}^n \mu_i = \mu, \quad (4.3.5)$$

$$\mathcal{E}E = \mathcal{E}\left(\sum_{i=1}^n E_i\right) = \sum_{i=1}^n \mathcal{E}E_i = 0. \quad (4.3.6)$$

The variance of the composite observed score is given by

$$\begin{aligned} \sigma_X^2 &= \sigma^2\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \sigma^2(Y_i) + \sum_{\substack{i=1 \\ i \neq j}}^n \sum_{j=1}^n \sigma(Y_i, Y_j) \\ &= \sum_{i=1}^n \sigma^2(Y_i) + \sum_{\substack{i=1 \\ i \neq j}}^n \sum_{j=1}^n \rho(Y_i, Y_j)\sigma(Y_i)\sigma(Y_j). \end{aligned} \quad (4.3.7)$$

Similarly we can see that the variance of the composite true score is given by

$$\sigma_T^2 = \sum_{i=1}^n \sigma^2(T_i) + \sum_{\substack{i=1 \\ i \neq j}}^n \sum_{j=1}^n \rho(T_i, T_j)\sigma(T_i)\sigma(T_j). \quad (4.3.8)$$

And, analogous to (4.2.6), the variance of the composite error score is

$$\sigma_E^2 = \sum_{i=1}^n \sigma^2(E_i) + \sum_{\substack{i=1 \\ i \neq j}}^n \sum_{j=1}^n \rho(E_i, E_j)\sigma(E_i)\sigma(E_j) = \sum_{i=1}^n \sigma^2(E_i), \quad (4.3.9)$$

since  $\rho(E_i, E_j) = 0$  for all  $i \neq j$ .

If the  $n$  measurements are parallel and therefore have equal variances, covariances, and error variances, then we obtain the following simplification:

$$\sigma_X^2 = n\sigma_Y^2[1 + (n - 1)\rho_{YY'}], \quad (4.3.10)$$

$$\sigma_T^2 = n^2\sigma^2(T_1), \quad (4.3.11)$$

$$\sigma_E^2 = n\sigma^2(E_1). \quad (4.3.12)$$

Equation (4.3.11) states that *if the length of a test is increased  $n$  times by adding parallel measurements, the composite true-score variance increases by the factor  $n^2$ .* In contrast, Eq. (4.3.12) states that *the variance of the composite error score increases only by the factor  $n$ , that is, in direct proportion with the length of the test.* The fact that true-score variance increases more rapidly than error variance is what makes it profitable to lengthen a test.

Often it is desirable to compare scores on tests of different length. To make this comparison and also legitimately to compare error variances and other properties of the tests, it is useful to turn our attention to relative scores. The variables  $X/n$ ,  $T/n$ , and  $E/n$  may be referred to as the *relative observed, true, and error scores.* They give the average observed, true, and error scores per measurement unit. The observed, true, and error variances of these relative scores are easily obtained:

$$\sigma^2(X/n) = (1/n)\sigma_Y^2[1 + (n - 1)\rho_{YY'}], \quad (4.3.13)$$

$$\sigma^2(T/n) = \sigma^2(T_1), \quad (4.3.14)$$

$$\sigma^2(E/n) = \sigma^2(E_1)/n. \quad (4.3.15)$$

We note here for future reference a formula of general use in dealing with weighted composites. For constants  $w_\alpha$ ,  $w_{\alpha'}$  and random variables  $Y_\alpha$ ,  $Y_{\alpha'}$ , the reader can prove the well-known result that

$$\sigma^2 \left[ \sum_{\alpha=1}^k w_\alpha Y_\alpha \right] = \sum_{\alpha=1}^k \sum_{\alpha'=1}^k w_\alpha w_{\alpha'} \sigma(Y_\alpha, Y_{\alpha'}), \quad (4.3.16)$$

where

$$\sigma(Y_\alpha, Y_{\alpha'}) = \sigma^2(Y_\alpha) \quad \text{if } Y_\alpha \equiv Y_{\alpha'}.$$

It should be noted that (4.3.16) may be rewritten as

$$\sigma^2 \left[ \sum_{\alpha=1}^k w_\alpha Y_\alpha \right] = \sum_{\alpha=1}^k w_\alpha^2 \sigma^2(Y_\alpha) + \sum_{\alpha=1}^k \sum_{\alpha'=1}^k \begin{matrix} k \\ \alpha \neq \alpha' \end{matrix} w_\alpha w_{\alpha'} \sigma(Y_\alpha, Y_{\alpha'}). \quad (4.3.17)$$

This cumbersome formula is easily used: Simply write the square of the term in brackets, replacing each product of the form  $y_\alpha, y_{\alpha'}$  by  $\sigma(Y_\alpha, Y_{\alpha'})$ . For example, to expand  $\sigma^2(aU - V)$ , consider that

$$(aU - V)^2 = a^2 U^2 - 2aUV + V^2$$

and write

$$\sigma^2(aU - V) = a^2\sigma_U^2 - 2a\sigma_{UV} + \sigma_V^2.$$

Now consider two sets of component random variables  $Z_\alpha$ ,  $\alpha = 1, 2, \dots, k$ , and  $Y_\beta$ ,  $\beta = 1, 2, \dots, l$ ; also consider weights  $a_\alpha$  and  $b_\beta$ . Then

$$\sigma \left[ \sum_{\alpha=1}^k a_\alpha Z_\alpha, \sum_{\beta=1}^l b_\beta Y_\beta \right] = \sum_{\alpha=1}^k \sum_{\beta=1}^l a_\alpha b_\beta \sigma(Z_\alpha, Y_\beta), \quad (4.3.18)$$

where

$$\sigma(Z_\alpha, Y_\beta) = \sigma^2(Z_\alpha) \quad \text{if } Z_\alpha \equiv Y_\beta. \quad (4.3.19)$$

This result states that the covariance between composite variables with weighted sums of components is a certain weighted sum of the covariances of the components. As before, the covariance of two sums can be written simply by multiplying the two sums and replacing terms like  $Z_\alpha Y_\beta$  by  $\sigma(Z_\alpha, Y_\beta)$ . For example, to expand  $\sigma(Y + Z, Y - Z)$ , consider the product  $Y^2 - Z^2$  and write  $\sigma(Y + Z, Y - Z) = \sigma_Y^2 - \sigma_Z^2$ .

#### 4.4 Coefficient $\alpha$ and the Reliability of Composite Measurements

In this section, we obtain a lower bound on the reliability of a composite test. We derive this result for the case of the composite as the sum of two component parts and then for the more general case. We show the general Spearman-Brown formula to be the special case of this result when the bound is attained. The inequalities obtained here were first derived by Guttman (1945).

**Theorem 4.4.1.** Let  $Y_1$  and  $Y_2$  be measurements with true scores  $T_1$  and  $T_2$ , and let  $X = Y_1 + Y_2$  be a composite measurement with true score  $T$ . Then

$$\rho_{XT}^2 \geq 2 \left[ 1 - \frac{\sigma^2(Y_1) + \sigma^2(Y_2)}{\sigma_X^2} \right]. \quad (4.4.1)$$

*Proof.* The proof of this theorem is based primarily on the result (3.5.4) that  $\sigma(Y_1, Y_2) = \sigma(T_1, T_2)$ , where  $Y_1$  and  $Y_2$  are observed scores with corresponding true scores  $T_1$  and  $T_2$ . Now

$$[\sigma(T_1) - \sigma(T_2)]^2 \geq 0;$$

hence

$$\sigma^2(T_1) + \sigma^2(T_2) \geq 2\sigma(T_1)\sigma(T_2).$$

The Cauchy-Schwartz inequality states that

$$\sigma(T_1)\sigma(T_2) \geq |\sigma(T_1, T_2)|, \quad (4.4.2)$$

that is, the correlation coefficient  $\rho(T_1, T_2) = \sigma(T_1, T_2)/\sigma(T_1)\sigma(T_2)$  is, in absolute value, less than or equal to one. Hence

$$\sigma^2(T_1) + \sigma^2(T_2) \geq 2|\sigma(T_1, T_2)| \geq 2\sigma(T_1, T_2).$$

Thus, since  $T = T_1 + T_2$ , by adding  $2\sigma(T_1, T_2)$  to the leftmost and rightmost members, we have

$$\sigma_T^2 = \sigma^2(T_1) + \sigma^2(T_2) + 2\sigma(T_1, T_2) \geq 4\sigma(T_1, T_2),$$

and hence, using (3.2.5) and (3.5.4), we have

$$\begin{aligned} \rho_{XT}^2 &= \frac{\sigma_T^2}{\sigma_X^2} \geq \frac{4\sigma(T_1, T_2)}{\sigma_X^2} = 2 \frac{2\sigma(Y_1, Y_2)}{\sigma_X^2} \\ &= 2 \left[ \frac{\sigma_X^2 - \sigma^2(Y_1) - \sigma^2(Y_2)}{\sigma_X^2} \right], \end{aligned}$$

from which the theorem follows immediately.  $\square$  An important test theory result is contained in the following corollary.

**Corollary 4.4.2.** If  $Y = Y_1$  and  $Y' = Y_2$  are parallel measurements, then (4.4.1) reduces to

$$\rho_{XT}^2 = \frac{2\rho_{YY'}}{1 + \rho_{YY'}}. \quad (4.4.3)$$

*Proof.* Note that  $\sigma^2(Y_1) = \sigma^2(Y_2)$  for parallel measurements. Using (4.2.7), we may easily put the right-hand member of (4.4.1) in the form (4.2.10), which has been previously derived as an equality.  $\square$

This result is known as the Spearman-Brown formula (4.2.10) for the reliability of a composite test of double length when the components are assumed to be parallel. The general conditions under which equality is attained in (4.4.1) are given in Corollary 4.4.3b.

In Exercise 4.16 the reader is asked to show that if  $Y_1$  and  $Y_2$  have equal variances and reliabilities but are not in fact parallel, then (4.4.3) gives a value smaller than the true reliability of  $Y_1 + Y_2$ .

The more general result corresponding to Theorem 4.4.1 is the following:

**Theorem 4.4.3.** Let  $Y_1, Y_2, \dots, Y_n$  be  $n$  measurements with true scores  $T_1, T_2, \dots, T_n$  and let  $X = Y_1 + Y_2 + \dots + Y_n$ . Then

$$\rho_{XX'} \equiv \rho^2(X, T) \geq \frac{n}{n-1} \left[ 1 - \frac{\sum \sigma^2(Y_i)}{\sigma_X^2} \right]. \quad (4.4.4)$$

*Proof.* As in the proof of (4.4.1), from the Cauchy-Schwartz inequality we have

$$\sigma^2(T_i) + \sigma^2(T_j) \geq 2\sigma(T_i, T_j).$$

Summing for  $i \neq j$ , we have

$$\sum_{i \neq j} [\sigma^2(T_i) + \sigma^2(T_j)] \geq 2 \sum_{i \neq j} \sigma(T_i, T_j). \quad (4.4.5)$$

We then note the mathematical identities

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^n [\sigma^2(T_i) + \sigma^2(T_j)] &= \sum_{i=1}^n [n\sigma^2(T_i) + \sum_{j=1}^n \sigma^2(T_j)] \\ &= n \sum_{i=1}^n \sigma^2(T_i) + n \sum_{j=1}^n \sigma^2(T_j) = 2n \sum_{i=1}^n \sigma^2(T_i)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^n [\sigma^2(T_i) + \sigma^2(T_j)] &= \sum_{i=j}^n \sum_{i \neq j} [\sigma^2(T_i) + \sigma^2(T_j)] + \sum_{i \neq j} \sum_{i \neq j} [\sigma^2(T_i) + \sigma^2(T_j)] \\ &= 2 \sum_{i=1}^n \sigma^2(T_i) + \sum_{i \neq j} \sum_{i \neq j} [\sigma^2(T_i) + \sigma^2(T_j)].\end{aligned}$$

Therefore the inequality (4.4.5) is *equivalent* to the inequality

$$\sum_{i=1}^n \sigma^2(T_i) \geq \frac{\sum_{i \neq j} \sigma(T_i, T_j)}{(n - 1)}. \quad (4.4.6)$$

Later we shall show that (4.4.6), as an equality, is the assumption made by Gulliksen (1950) in his derivation of coefficient  $\alpha$  as the reliability of a test. Now

$$\sigma^2(T) = \sigma^2(\sum T_i) = \sum \sigma^2(T_i) + \sum_{i \neq j} \sum \sigma(T_i, T_j).$$

Substituting from (4.4.6), we have

$$\sigma^2(T) \geq \frac{\sum_{i \neq j} \sigma(T_i, T_j)}{n - 1} + \sum_{i \neq j} \sum \sigma(T_i, T_j),$$

and, combining terms on the right,

$$\sigma^2(T) \geq \frac{n}{n - 1} \sum_{i \neq j} \sum \sigma(T_i, T_j). \quad (4.4.7)$$

To complete the proof, note that

$$\sigma^2(X) - \sum \sigma^2(Y_i) = \sum_{i \neq j} \sum \sigma(Y_i, Y_j) = \sum_{i \neq j} \sum \sigma(T_i, T_j).$$

Now substituting into (4.4.7) and dividing by  $\sigma^2(X)$ , we have

$$\rho^2(X, T) \equiv \frac{\sigma^2(T)}{\sigma_X^2} \geq \frac{n}{n - 1} \left[ 1 - \frac{\sum \sigma^2(Y_i)}{\sigma_X^2} \right]. \quad \square$$

The rightmost member of (4.4.4),

$$\alpha \equiv \frac{n}{n - 1} \left[ 1 - \frac{\sum \sigma^2(Y_i)}{\sigma_X^2} \right], \quad (4.4.8)$$

has been termed *coefficient  $\alpha$*  by Cronbach (1951). Thus coefficient  $\alpha$ , a quantity that can be computed from the results of one test administration, gives a lower bound on the reliability of the test. An important special case is contained in

**Corollary 4.4.3a.** If  $Y_1, Y_2, \dots, Y_n$  are  $n$  parallel measurements, then, as a generalization of (4.2.10), we may obtain

$$\rho_{XX'} = \frac{n\rho_{YY'}}{1 + (n - 1)\rho_{YY'}}. \quad (4.4.9)$$

*Proof.* If  $Y_1, Y_2, \dots, Y_n$  are parallel measurements, and if we note that  $\sigma^2(Y_1) = \sigma^2(Y_2) = \dots = \sigma^2(Y_n)$  and use (4.3.10), we can put the right-hand member of (4.4.4) in the form (4.4.9). Actually, equality obtains in (4.4.2) and (4.4.4) if the  $Y_i$  are  $\tau$ -equivalent measurements.  $\square$

This result is known as the (*general*) Spearman-Brown formula for the reliability of a composite test having parallel components. Thus the reliability of a composite test having parallel components is given by coefficient  $\alpha$ . (A more direct derivation of the result is outlined in Exercise 4.8).

The general necessary and sufficient condition for equality in the Cauchy-Schwartz inequality is, in our context, the condition that  $T_i$  be a linear function of  $T_j$ , which is implied by the definition of  $\tau$ -equivalent measurements. In the proof of (4.4.4), the only other inequalities used were  $[\sigma(T_i) - \sigma(T_j)]^2 \geq 0$  and  $|\sigma(T_i, T_j)| \geq \sigma(T_i, T_j)$ . Given that  $T_i = a_{ij} + b_{ij}T_j$ , the first of these becomes an equality if and only if  $b_{ij} = \pm 1$ , and the second, if and only if  $|b_{ij}| = b_{ij}$ . Together, these imply  $b_{ij} = 1$ . Thus we have

**Corollary 4.4.3b.** The necessary and sufficient condition for (4.4.4) to hold as an equality is that  $T_i = a_{ij} + T_j$  for all  $i, j$ .

Formula (4.4.4) has been the subject of a rather considerable body of research, both as an inequality and as an equality, and in its given form and in certain forms for special cases [see Cronbach (1951) primarily, but also Kuder and Richardson (1937); Rulon (1939), Dressel (1940), Hoyt (1941), Guttman (1945), Cureton (1958), Lyerly (1958), and Novick and Lewis (1967)].

Let us give a practical application of the preceding theory. A preliminary form of the Educational Testing Service *Test of English as a Foreign Language* was administered on May 2, 1965 to 1416 foreign students seeking admission to U.S. Colleges. We have used results from four parts of this test in preparing the following example. The reader should note that we have treated sample quantities as if they were population parameters; however, since the sample is a relatively large one, there is little danger in doing so. Approximate reliabilities of the four parts were 0.889, 0.945, 0.933, and 0.900, respectively. These were obtained by computing coefficient  $\alpha$  for each subtest. This procedure is justifiable because the items within each scale have been selected to measure the same thing and hence tend to have similar true scores, and also because the

item length  $n$  of each subtest is large. The variance-covariance matrix was

94.7	87.3	63.9	58.4
	212.0	138.7	128.2
		160.5	109.8
			115.8

The reliability of the composite test (simple sum of components) was determined (using 4.7.2) to be 0.950. From (4.4.4), coefficient  $\alpha$  for the composite test (simple sum) is

$$\begin{aligned}\alpha &= \frac{4}{3} \left[ 1 - \frac{94.7 + 212.0 + 160.5 + 115.8}{94.7 + \dots + 115.8 + 2(87.3 + 138.7 + 109.8 + 63.9 + 128.2 + 58.4)} \right] \\ &= \frac{4}{3} \left[ 1 - \frac{582.97}{1755.33} \right] = 0.891.\end{aligned}$$

In this case,  $\alpha$  does not provide a particularly good lower bound on reliability of the total test because the components (the subtests) are quite heterogeneous and  $n$  is small (see Exercise 4.18). However, as indicated in Chapter 6, under many circumstances coefficient  $\alpha$  gives a very good lower bound.

Kuder and Richardson had originally discussed special cases of this coefficient's arising when the individual components are binary items, i.e., random variables  $u_i$  taking values zero and unity with probabilities  $1 - p_i$  and  $p_i$ . Under these conditions, coefficient  $\alpha$  reduces to the *Kuder-Richardson formula 20* (KR20):

$$\alpha_{(20)} = \frac{n}{n-1} \left( 1 - \frac{\sum_{i=1}^n p_i q_i}{\sigma_X^2} \right). \quad (4.4.10)$$

If the items have identical  $p_i$ -values, then (4.4.8) reduces to the *Kuder-Richardson formula 21* (KR21):

$$\alpha_{(21)} = \frac{n}{n-1} \left( 1 - \frac{n \bar{p} \bar{q}}{\sigma_X^2} \right), \quad (4.4.11)$$

where  $\bar{p} = \sum p_i/n$ ,  $\bar{q} = \sum_{i=1}^n (1 - p_i)$ . It is easily seen that  $\alpha_{(20)} \geq \alpha_{(21)}$ , with equality obtaining if and only if all the  $p_i$  are equal. We leave the proof of these results as an exercise for the reader. The two formulas were given as reliabilities by Kuder and Richardson (1937), and derived under a rather complicated assumption. The Kuder-Richardson formula 21 is also a lower bound on reliability, although not so good a bound as formula 20 if in fact the  $p$ -values are not identical. It does have an advantage of computational simplicity. How-

ever, except for purposes of rather informal analysis, it is seldom good practice to use this formula in place of the more appropriate formula 20. While tests may be relatively "homogeneous" in content, they are seldom "homogeneous" in item mean observed score (however, see Exercise 4.17.) For a clear picture of the relationship between KR20 and KR21, see Exercise 4.21.

A standard derivation of coefficient  $\alpha$  as an equality (Gulliksen, 1950) proceeds by considering two  $n$ -item tests that are parallel item for item and then introducing the assumption that the covariance of an item in one test with the parallel (matched) item in the other test is equal, on the average, to the covariance between pairs of items within a single test. Gulliksen stated that this is "the simplest and most direct assumption" to make in order to derive coefficient  $\alpha$ .

Although this assumption is certainly direct, its justification is not immediately apparent. We shall show, however, that this assumption is equivalent to the condition of Corollary 4.4.3b. If  $\rho_{ij}$  and  $\sigma_i$  denote observed item correlations and standard deviations, the Gulliksen assumption may be written

$$\sum_{i=1}^n \rho_{ii'}\sigma_i^2 = \frac{\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j}{n-1}, \quad (4.4.12)$$

where  $i$  and  $i'$  are the matched items on the two tests, and  $i$  and  $j$  are items on the same test, there being  $n$  covariances between matched items and  $n(n-1)$  covariances within each form. But

$$\rho_{ii'}\sigma_i^2 = \sigma(T_i, T_{i'}) = \sigma^2(T_i) \quad \text{and} \quad \rho_{ij}\sigma_i\sigma_j = \sigma(Y_i, Y_j) = \sigma(T_i, T_j).$$

Hence (4.4.12) may be written as

$$\sum_{i=1}^n \sigma^2(T_i) = \frac{\sum_{i \neq j} \sigma(T_i, T_j)}{n-1}.$$

The equivalence of this assumption with the assumption  $T_i = a_{ij} + T$  is contained in the proof of Corollary 4.4.3a (see Eq. 4.4.6). Hence we have

#### Corollary 4.4.4

$$\sum_{i=1}^n \rho_{ii'}\sigma_i^2 = \frac{\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j}{n-1} \quad \text{if and only if } T_i = a_{ij} + T_j.$$

That assumption (4.4.12) is merely a complicated form of the condition of Corollary 4.4.3b should in part be obvious on an intuitive basis, since for fixed true-score variances the covariance between two items is a maximum when the items are parallel (or essentially  $\tau$ -equivalent). But assumption (4.4.12) states that the average of covariances of arbitrary items is equal to the average of parallel items. This, of course, can be true only if the arbitrary items are in fact essentially  $\tau$ -equivalent.

Coefficient  $\alpha$  has one further important property, which was established by Cronbach (1951). Suppose we have a test of  $2n$  items. The test may be divided into split halves  $\frac{1}{2}(2n)!/(n!)^2$  ways. For each of these divisions, we may assume  $n = 2$  in the right-hand member of (4.4.8) and compute coefficient  $\alpha$  in terms of the two half-tests forming the composite, denoting it by  $\alpha_2$ . We may also compute the coefficient  $\alpha$  for the  $2n$ -item test and denote it by  $\alpha_{2n}$ . We then have

**Theorem 4.4.5**

$$\mathcal{E}^*\alpha_2 = \alpha_{2n}. \quad (4.4.13)$$

The expectation  $\mathcal{E}^*$  of  $\alpha_2$  over the possible splits is equal to  $\alpha_{2n}$ . This result may be interpreted in two ways: First it states that if we compute all  $\frac{1}{2}(2n)!/(n!)^2$  possible values of  $\alpha_2$  and take their average, we shall obtain a value equal to  $\alpha_{2n}$ . Also it states that if items are assigned to the two half-tests randomly, then the expected value of  $\alpha_2$  is  $\alpha_{2n}$ .

*Proof*

$$\begin{aligned} \frac{n}{n-1} \left[ 1 - \frac{\sum \sigma^2(Y_i)}{\sigma_X^2} \right] &= \frac{n}{n-1} \left[ \frac{\sum_{i \neq j} \sigma(Y_i, Y_j)}{\sigma_X^2} \right] \\ &= \frac{1}{n(n-1)} \left[ \frac{\sum_{i \neq j} \sigma(Y_i, Y_j)}{n^{-2}\sigma_X^2} \right]. \end{aligned} \quad (4.4.14)$$

Let  $Z_1$  and  $Z_2$  be arbitrary split halves. With no loss of generality, we take

$$Z_1 = \sum_{i=1}^n Y_i \quad \text{and} \quad Z_2 = \sum_{i=n+1}^{2n} Y_i,$$

where we assume the test to have  $2n$  items. For this split, the coefficient  $\alpha$ , which we denote by  $\alpha_2$ , is

$$\alpha_2 = 2 \left[ 1 - \frac{\sigma^2(Z_1) + \sigma^2(Z_2)}{\sigma_X^2} \right] = \frac{4\sigma(Z_1, Z_2)}{\sigma_X^2}.$$

Expanding  $\sigma(Z_1, Z_2)$  as the covariance of two composites, we have

$$\alpha_2 = \frac{4}{\sigma_X^2} \sum_{i=1}^n \sum_{j=n+1}^{2n} \sigma(Y_i, Y_j).$$

Taking the expected value over all splits (each of which is assumed to be equally likely), we have

$$\mathcal{E}^*\alpha_2 = \frac{4}{\sigma_X^2} \sum_{i=1}^n \sum_{j=n+1}^{2n} \mathcal{E}\sigma(Y_i, Y_j).$$

For an arbitrary split, all distinct pairs of components have an equal chance

of being assigned to the labels  $Y_i$  and  $Y_j$  for  $i = 1, 2, 3, \dots, n$  and  $j = n+1, \dots, 2n$ . Hence all terms in the double sum are equal and, moreover, equal to the average covariance between distinct components of the test. Since there are  $n^2$  terms in the sum, we may therefore write  $\mathcal{E}^*\alpha_2$  as

$$\mathcal{E}^*\alpha_2 = \frac{4n^2}{\sigma_X^2} \sum_{i \neq j}^{2n} \frac{\sigma(Y_i, Y_j)}{2n(2n-1)},$$

where we have singled out a particular labeling of the components that allows each of the subscripts  $i$  and  $j$  to range from 1 to  $2n$ , for  $i \neq j$ . The general formula for  $\alpha_{2n}$  is given in (4.4.14). Using this, we may write coefficient  $\alpha$  for a test of  $2n$  items as

$$\alpha_{2n} = \frac{1}{2n(2n-1)} \frac{\sum_{i \neq j}^{2n} \sigma(Y_i, Y_j)}{(1/4n^2)\sigma_X^2}.$$

Hence  $\mathcal{E}\alpha_2 = \alpha_{2n}$ .  $\square$  Also we have

**Corollary 4.4.6.** If the  $2n$  items of the test are  $\tau$ -equivalent, then

$$\alpha_2 \equiv \alpha_{2n}.$$

We leave the proof of the corollary as an exercise for the reader. For an application of the theorem see Exercise 4.12.

Typically the values of  $\alpha_2$  differ for the different possible splits of the test. From Theorem 4.4.3 we know that each of these coefficients is a lower bound, and hence we may conclude that, unless they are all equal, at least one of them is a better lower bound than  $\alpha_{2n}$ . Indeed, since the formula for coefficient  $\alpha$  does not assume that the components are split *halves*, if the various coefficients  $\alpha$  could be estimated very accurately, it would be possible to consider all possible divisions of the test into components and to take the maximum value of the coefficient  $\alpha$  thus obtained as the “best” lower bound on reliability.

The original study of lower bounds on reliability is due to Guttman (1945), who formulated most of the important bounds now available. However, a more readable and integrated treatment has been given by Koutsopoulos (1964). These papers include descriptions of a number of other lower bounds on reliability, each of which is obtainable on the basis of a single administration of the total test. One of these bounds, called  $\lambda_3$  by Guttman, is of interest because it is always at least as good as coefficient  $\alpha$ ; typically it is modestly better and occasionally it is substantially better. We give this bound as

**Theorem 4.4.7**

$$\rho^2(X, T) \geq 1 - \frac{\sum \sigma^2(Y_i)}{\sigma_X^2} + \frac{\sqrt{(n/n-1) \sum_{i \neq j} \sigma^2(Y_i, Y_j)}}{\sigma_X^2}. \quad (4.4.15)$$

*Proof.* Following Koutsopoulos, we note that

$$\sigma^4(T_i) + \sigma^4(T_j) \geq 2\sigma^2(T_i)\sigma^2(T_j) \geq 2\sigma^2(T_i, T_j) = 2\sigma^2(Y_i, Y_j).$$

Next, following the method of proof of the previous theorem, we obtain

$$\sum_{i=1}^n \sigma^4(T_i) \geq \frac{1}{n-1} \sum_{i \neq j} \sum \sigma^2(Y_i, Y_j).$$

Now

$$\left[ \sum_i \sigma^2(T_i) \right]^2 - \sum_i \sigma^4(T_i) = \sum_{i \neq j} \sum \sigma^2(T_i)\sigma^2(T_i) \geq \sum_{i \neq j} \sum \sigma^2(Y_i, Y_j),$$

and hence

$$\left[ \sum_i \sigma^2(T_i) \right]^2 \geq \sum_i \sigma^4(T_i) + \sum_{i \neq j} \sum \sigma^2(Y_i, Y_j).$$

The result follows on substitution for

$$\sum_{i=1}^n \sigma^4(T_i). \quad \square$$

Guttman's  $\lambda_3$  is particularly useful when some of the test items have negative intercorrelations. In such a case, coefficient  $\alpha$  can be negative and therefore useless as a lower bound on reliability;  $\lambda_3$ , however, is always positive.

#### 4.5 The Internal Structure of Tests

The question of test homogeneity is one which has been discussed at length in the test theory literature. Unfortunately there is no general agreement as to just what this term should mean and how homogeneity should be measured. We shall not attempt a complete explication in this book.

On the other hand, our work on coefficient  $\alpha$  provides a base for a definition of the homogeneous test that is perfectly satisfactory for our present purposes. Since  $\alpha$  is equal to the reliability of a test if and only if all components are essentially  $\tau$ -equivalent, it seems natural to define a homogeneous test as one whose components have this property. Thus a homogeneous test is one whose components all "measure the same thing" in their true-score components. It is important to note that the scalings of the measurements may differ by an additive constant. A similar approach has been taken by Meredith (1965), although his treatment is set in a somewhat broader context than the development in this chapter. In this broader context,  $X_1$  and  $X_2$  are homogeneous if  $T_1 \equiv \varphi(T_2)$ , where  $\varphi$  is a strictly increasing function. Thus  $T_1$  and  $T_2$  "measure the same thing" but in scalings which are not necessarily linearly related. The work presented in this chapter would suggest that one useful explication of the concept of homogeneity might be made in terms of the variance-covariance matrix of true scores.

#### 4.6 Expectations, Variances, and Covariances of Weighted Composites

We can further extend the concept of a composite measurement by considering component measurements  $(Y_{ga}, \tau_{ga}, E_{ga})$  for a fixed person  $a$ , taking values  $(y_{ga}, \tau_{ga}, e_{ga})$ ,  $g = 1, 2, \dots, n$ , and a weighted composite measurement  $(X_a, \tau_a, E_a)$ , taking values  $(x_a, \tau_a, e_a)$  and defined by

$$X_a \equiv \sum_{g=1}^n w_g Y_{ga}. \quad (4.6.1)$$

Then

$$\mathcal{E}X_a = \sum_{g=1}^n w_g \mathcal{E}Y_{ga} \quad (4.6.2)$$

and

$$\sigma^2(X_a) = \sigma^2 \left( \sum_{g=1}^n w_g Y_{ga} \right) = \sum_{g=1}^n w_g^2 \sigma^2(Y_{ga}). \quad (4.6.2a)$$

There are no covariance terms, since  $Y_{ia}$  and  $Y_{ja}$  are assumed to be experimentally independent for  $i \neq j$ .

For a randomly selected person, the component measurements are  $(Y_g, T_g, E_g)$ , taking values  $(y_g, \tau_g, e_g)$ ,  $g = 1, 2, 3, \dots, n$ , and the composite measurement is  $(X, T, E)$ , taking values  $(x, \tau, e)$  and defined by

$$X = \sum_{g=1}^n w_g Y_g. \quad (4.6.3)$$

Then, taking expectations over people, we have

$$\mathcal{E}X = \sum_{g=1}^n w_g \mathcal{E}Y_g \quad (4.6.4)$$

$$\sigma^2(X) = \sigma^2 \left( \sum_{g=1}^n w_g Y_g \right) = \sum_{g=1}^n \sum_{g'=1}^n w_g w_{g'} \sigma(Y_g, Y_{g'}), \quad (4.6.5)$$

where  $\sigma(Y_g, Y_{g'}) = \sigma^2(Y_g)$  when  $g = g'$ .

Equation (4.6.1) and hence Eq. (4.6.3) can be simplified with no loss of generality by absorbing the scoring weight into the component scores, i.e., by defining

$$Y'_{ga} = w_g Y_{ga} \quad (4.6.6)$$

and then dropping the prime, so that

$$X_a = \sum_{i=1}^n Y_{ga}. \quad (4.6.7)$$

In Chapters 7 through 9, it will be convenient to assume that this rescaling has been accomplished unless otherwise specified.

Most cognitive and many other kinds of tests are composed of items that in practice are scored dichotomously. Much of the treatment in this book is limited to this special case, partly because it is simpler to understand, but also because much of the theory is not otherwise available.

It is usually convenient to consider that a dichotomously scored item is scored either 0 or 1, in which case we call it a *binary item*. Formula (4.6.7) includes the special case of *binary items*. It is also convenient to denote the binary variable by  $U_{ga}$ , instead of by  $Y_{ga}$ , and to say that the response of examinee  $a$  to item  $g$  is a *right answer* ( $U_{ga} = 1$ ) or a *wrong answer* ( $U_{ga} = 0$ ). The test score  $x_a$  is then the number of right answers for examinee  $a$ . We denote the proportion of right answers on item  $g$  in the population of examinees by

$$\pi_g \equiv \mathbb{E}_a U_{ga} \quad (4.6.8)$$

and call this proportion *item difficulty*. This terminology is used even for items such as attitude items for which the terms “right”, “wrong”, and “difficulty” have no natural meaning, and also for nonbinary items. For binary random variables, the variance of the item can be given in terms of the item difficulty:

$$\sigma_g^2 = \pi_g - \pi_g^2. \quad (4.6.9)$$

#### 4.7 The Correlation between Two Composite Measurements

The general formula for the correlation between two composite variables may be obtained directly from (4.6.5) and (4.3.18). We have

$$\begin{aligned} & \rho \left( \sum_{\alpha=1}^k a_{\alpha} Z_{\alpha}, \sum_{\beta=1}^l b_{\beta} Y_{\beta} \right) \\ &= \frac{\sigma \left( \sum_{\alpha=1}^k a_{\alpha} Z_{\alpha}, \sum_{\beta=1}^l b_{\beta} Y_{\beta} \right)}{\left[ \sigma^2 \left( \sum_{\alpha=1}^k a_{\alpha} Z_{\alpha} \right) \right]^{1/2} \left[ \sigma^2 \left( \sum_{\beta=1}^l b_{\beta} Y_{\beta} \right) \right]^{1/2}} \\ &= \frac{\sum_{\alpha=1}^k \sum_{\beta=1}^l a_{\alpha} b_{\beta} \sigma(Z_{\alpha}, Y_{\beta})}{\left[ \sum_{\alpha=1}^k \sum_{\alpha'=1}^k a_{\alpha} a_{\alpha'} \sigma(Z_{\alpha}, Z_{\alpha'}) \right]^{1/2} \left[ \sum_{\beta=1}^l \sum_{\beta'=1}^l b_{\beta} b_{\beta'} \sigma(Y_{\beta}, Y_{\beta'}) \right]^{1/2}}. \quad (4.7.1) \end{aligned}$$

This formula has one very important application. Suppose a test yields  $k$  separate scores. If, in application, some or all of these scores are used in a weighted composite as a predictor of some criterion, then the reliability of the individual scores is of little immediate relevance. What is relevant is the reliability of the weighted composite to be used for prediction purposes. Thus a

test publisher who is to allot 150 items among ten types may “do better” if he produces ten subtests that are each only moderately reliable than if he constructs a single 150-item scale that is highly reliable. An appropriate linear composite of the ten 15-item subtests would probably be as reliable as the 150-item homogeneous test, and the composite can probably be made a better predictor of important criteria.

Formula (4.7.1) can be specialized to the case in which  $a_i = b_i$  and  $Z_i$  is parallel to  $Y_i$  for  $i = 1, 2, \dots, k$ . The resulting expression is just the reliability of the composite measurement  $X = \sum Y_i$ . Letting  $X$  have true score  $T$ , we have

$$\rho^2(X, T) = \frac{\sum_{\alpha=1}^k \sigma^2(Y_\alpha) \rho(Y_\alpha, Z_\alpha) + \sum_{\alpha \neq \alpha'} \sigma(Y_\alpha, Y_{\alpha'})}{\sum_{\alpha=1}^k \sigma^2(Y_\alpha) + \sum_{\alpha \neq \alpha'} \sigma(Y_\alpha, Y_{\alpha'})}. \quad (4.7.2)$$

The proof of this result is left as an exercise for the reader.

The special case of (4.7.1) in which the components are parallel and equal weights are used is of particular importance. This result has been given as Eq. (4.4.9). A second special case of particular interest is that in which one of the composites has but a single component. For example, suppose  $k = 1$ . Then

$$\rho(a_1 Z_1, \sum b_\beta Y_\beta) = \frac{\sum_{\beta=1}^l b_\beta \sigma(Z_1, Y_\beta)}{\sigma(Z_1) \left[ \sum_{\beta=1}^l \sum_{\beta'=1}^l b_\beta b_{\beta'} \sigma(Y_\beta, Y_{\beta'}) \right]^{1/2}}. \quad (4.7.3)$$

Suppose we are given  $n$  components divided into two groups of sizes  $m$  and  $n - m$ , where  $n > m$ . Suppose that composite variables

$$X_m = \sum_{\alpha=1}^m w_\alpha Y_\alpha \quad \text{and} \quad X_n = \sum_{\alpha=1}^m w_\alpha Y_\alpha + \sum_{\alpha=m+1}^n w_\alpha Y_\alpha$$

are defined. The composite  $X_m$  is thus a part of the composite  $X_n$ . Then the correlation between  $X_m$  and  $X_n$  is given by

$$\rho(X_m, X_n) = \frac{\sum_{\alpha=1}^m \sum_{\alpha'=1}^m w_\alpha w_{\alpha'} \sigma(Y_\alpha, Y_{\alpha'}) + \sum_{\alpha=1}^m \sum_{\alpha'=m+1}^n w_\alpha w_{\alpha'} \sigma(Y_\alpha, Y_{\alpha'})}{\left[ \sum_{\alpha=1}^m \sum_{\alpha'=1}^m w_\alpha w_{\alpha'} \sigma(Y_\alpha, Y_{\alpha'}) \right]^{1/2} \left[ \sum_{\alpha=1}^n \sum_{\alpha'=1}^n w_\alpha w_{\alpha'} \sigma(Y_\alpha, Y_{\alpha'}) \right]} . \quad (4.7.4)$$

This describes the (“spurious”) correlation between a total test score and a subtest score (“spurious” because the subtest score and its error are part of the total test score).

In the special case where all  $w_\alpha = 1$ , (4.7.4) reduces to

$$\rho(X_m, X_n) = \frac{\sum_{\alpha=1}^m \sum_{\alpha'=1}^m \sigma(Y_\alpha, Y_{\alpha'}) + \sum_{\alpha=1}^m \sum_{\alpha'=m+1}^n \sigma(Y_\alpha, Y_{\alpha'})}{\left[ \sum_{\alpha=1}^m \sum_{\alpha'=1}^m \sigma(Y_\alpha, Y_{\alpha'}) \right]^{1/2} \left[ \sum_{\alpha=1}^n \sum_{\alpha'=1}^n \sigma(Y_\alpha, Y_{\alpha'}) \right]^{1/2}}. \quad (4.7.5)$$

For  $m = 1$ , (4.7.4) becomes

$$\rho(X_1, X_n) = \frac{w_1 \sigma^2(Y_1) + \sum_{\alpha=2}^n w_\alpha \sigma(Y_1, Y_\alpha)}{\sigma(Y_1) \left[ \sum_{\alpha=1}^n \sum_{\alpha'=1}^n w_\alpha w_{\alpha'} \sigma(Y_\alpha, Y_{\alpha'}) \right]^{1/2}}. \quad (4.7.6)$$

## Exercises

- 4.1. Using the notation but not the results of Theorem 4.4.1, show that

$$2 \left( 1 - \frac{\sigma^2(Y_1) + \sigma^2(Y_2)}{\sigma_X^2} \right) = \frac{\sigma^2(T_1 + T_2) - \sigma^2(T_1 - T_2)}{\sigma_X^2} \leq \frac{\sigma_T^2}{\sigma_X^2}$$

by using the standard result  $\sigma(Y_1, Y_2) = \sigma(T_1, T_2)$ . Show how this result pinpoints the conditions under which coefficient  $\alpha$  is equal to the reliability of the composite.

- 4.2. Let  $X = Y_1 + Y_2 + \dots + Y_n$  and  $X' = Y'_1 + Y'_2 + \dots + Y'_n$  be parallel composite measurements. Do not assume that pairs  $Y_g, Y_{g'}$  are either parallel or comparable. Let  $\sigma_{gh}$  be the covariance of  $Y_g$  with  $Y_h$ ,  $\sigma_{gh'}$  the covariance of  $Y_g$  with  $Y_{h'}$ , and  $\sigma_{gg} = \sigma_g^2$ . Show that a sufficient condition that

$$\rho_{XX'} = \frac{n}{n-1} \left( 1 - \frac{\sum \sigma_g^2}{\sigma_X^2} \right)$$

is that

$$\frac{1}{n^2} \sum_g \sum_{h'} \sigma_{gh'} = \frac{1}{n(n-1)} \sum \sum_{g \neq h} \sigma_{gh}.$$

- 4.3. Using Corollary 4.4.3a, show that the components of  $X$  and  $X'$  in Exercise 4.2 are essentially  $\tau$ -equivalent.

- 4.4. Let  $X = \sum_{g=1}^n Y_g$  and  $X' = \sum_{g'=1}^n Y'_{g'}$  be composite measurements having  $\tau$ -equivalent components. Suppose that each  $Y_g$  is parallel to the corresponding  $Y'_{g'}$ . Show that the reliability of  $X$  (or of  $X'$ ) is given by

$$\rho_{XT}^2 = \frac{n \sum_{g=g'} \sum \sigma^2(Y_g) \rho(Y_g, Y'_{g'})}{\sum_{g=1}^n \sigma^2(Y_g) + (n-1) \sum_{g=g'} \sum \sigma^2(Y_g) \rho(Y_g, Y'_{g'})}.$$

- 4.5. Let  $X = Y_1 + Y_2 + \dots + Y_n$  be a composite measurement. Let  $\rho_{XX'}$  be the reliability of  $X$ , and  $\rho_{ii'}$ ,  $i = 1, 2, \dots, n$ , be the reliabilities of the components. Let  $\sigma_i^2$  be the variance of the  $i$ th component and let  $\sigma_X^2$  be the variance of  $X$ . Show that

$$\rho_{XX'} = 1 - \frac{\sum \sigma_i^2 - \sum \rho_{ii'} \sigma_i^2}{\sigma_X^2}.$$

- 4.6. Let

$$X = \sum_{i=1}^n Y_i \quad \text{and} \quad X' = \sum_{i=1}^n Y'_i$$

each be a composite measurement with components not necessarily  $\tau$ -equivalent. However, assume that  $Y_i$  and  $Y'_i$  are parallel for each  $i$ . Let  $\alpha_n$  be coefficient  $\alpha$  for measurement  $X$  (or equivalently for measurement  $X'$ ). Let  $\alpha_{2n}$ , with obvious interpretation, be coefficient  $\alpha$  for the measurement  $(X + X')$ . Express  $\alpha_{2n}$  as a function of  $n$ ,  $\alpha_n$ , the item variances, and the item reliabilities. Show that when each composite has parallel components this formula reduces to the Spearman-Brown formula for a test of double length.

- 4.7. Jackson and Ferguson derived coefficient  $\alpha$  as the reliability of the composite test under the assumption that

$$\frac{\sum_g \sum_{h'} \sigma_{gh'}}{n^2} = \frac{\sum_{g \neq h} \sigma_{gh}}{n(n-1)},$$

where  $g, h, \dots$  and  $g', h', \dots$  are each sets on  $n$  components of two parallel composites. (It is not *assumed* that the elements of  $g$  or of  $h$  are essentially  $\tau$ -equivalent.) Show that this assumption is equivalent to the assumption that

$$T_g = \alpha_{gh} + T_h \quad \text{and} \quad T_{g'} = \alpha_{g'h'} + T_{h'},$$

that is, that the Jackson-Ferguson assumption is equivalent to the assumption of Corollary 4.4.3b. Show that therefore

- a) if we have a composite measurement which satisfies the Jackson-Ferguson assumption with respect to *any* parallel composite measurement (it not being assumed that the components are pairwise parallel), then the Gulliksen assumption is satisfied for the first measurement and *any* third composite measurement having components which are pairwise parallel to those of the first measurement; and
- b) if we have a composite measurement which satisfies the Gulliksen assumption with respect to *any* second measurement having components pairwise parallel to those of the first measurement, then the Jackson-Ferguson assumption is satisfied for the first measurement and *any* third composite measurement which is parallel to it (even though the components of these two composites are not pairwise parallel).

- 4.8. Derive (4.4.9) directly from (4.3.10), (4.3.11), and (3.3.7).

- 4.9. If  $X_1, X_2, \dots, X_g, \dots, X_n$  are parallel measurements on which the common true-score variable  $T$  takes the value  $\tau_a$  for person  $a$ , show that

$$\lim_{n \rightarrow \infty} \frac{\sum_{g=1}^n x_{ga}}{n} = \tau_{ga}$$

converges in probability to  $\tau_a$  as  $n \rightarrow \infty$ .

- 4.10. Derive Eq. (4.7.6).
- 4.11. For the example of Section 4.2, compute each of the values of  $\alpha_2$  and show that their average is  $\alpha$ .
- 4.12. For this same example, determine the part-whole correlation of each part.
- 4.13. Show that if a test is lengthened by adding parallel forms, the limiting value of coefficient  $\alpha$  as  $n \rightarrow \infty$  is unity, except in the exceptional case in which all the components are uncorrelated and in which the limiting value is therefore zero.
- 4.14. Show that Guttman's  $\lambda_3$  is always at least as good as coefficient  $\alpha$ .
- 4.15. Show that (4.2.10) holds if one assumes only that

$$\sigma^2(Y_1) = \sigma^2(Y_2) = \sigma^2(Y_3) = \sigma^2(Y_4)$$

and that

$$\sigma(Y_1, Y_4) = \sigma(Y_2, Y_4) = \sigma(Y_1, Y_3) = \sigma(Y_2, Y_3).$$

Explain how this result differs from (4.2.11).

- 4.16. Show that the Spearman-Brown formula for the reliability of a composite test of double length gives a value smaller than the reliability coefficient if the halves are not in fact parallel, but have equal means, variances, and reliabilities.
- 4.17. Show that binary items can be homogeneous in some populations even though they have different difficulty levels. Explain why this almost never occurs in practice.
- 4.18. Suppose parallel forms are constructed for each of the four subtests of the *Test of English as a Foreign Language*. Compute the reliability of the test consisting of these eight subtests. Also compute the new coefficient  $\alpha$ .
- 4.19. Show that if all  $p_i$  are not equal, then KR21 is less than KR20.
- 4.20. Suppose that a test is composed of binary items. If  $\alpha$  is equal to the reliability of the test and if the distribution of people on each item has  $p$  values  $0 \leq p \leq 1$ , show that coefficient  $\alpha$  reduces to KR21 and hence KR21 is the reliability of the test.
- 4.21. Let  $\sigma_p^2$  be the variance of the  $p_i$  in (4.4.10). Show that KR20 can be written as

$$\alpha_{20} = \frac{n}{n-1} \left( 1 + \frac{n\sigma_p^2 - n\bar{p}\bar{q}}{\sigma_X^2} \right),$$

which differs from KR21 only by the addition of the term  $n\sigma_p^2$  in the numerator.

### References and Selected Readings

- CRONBACH, L. J., Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297-334.
- CRONBACH, L. J., P. SCHÖNEMANN, and D. MCKIE, Alpha coefficients for stratified parallel tests. *Educational and Psychological Measurement*, 1965, **25**, 291-312.
- CURETON, E. E., The definition and estimation of test reliability. *Educational and Psychological Measurement*, 1958, **18**, 715-738.
- DRESSEL, P. L., Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 1940, **5**, 305-310.
- GHISELLI, E. E., *Theory of psychological measurement*. New York: McGraw-Hill, 1964, Chapter 7.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.
- GUTTMAN, L., A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255-282.
- HOYT, C., Test reliability estimated by analysis of variance. *Psychometrika*, 1941, **6**, 153-160.
- JACKSON, R. W. B., and G. A. FERGUSON, Studies on the reliability of tests. Bulletin 12. Toronto: Department of Educational Research, University of Toronto, 1941.
- KOUTSOPoulos, C. J., The mathematical foundations of classical test theory: an axiomatic approach, I. *Research Bulletin 62-17*. Princeton, N.J.: Educational Testing Service, 1962.
- KOUTSOPoulos, C. J., The mathematical foundations of classical test theory: an axiomatic approach, II. *Research Memorandum 64-3*. Princeton, N.J.: Educational Testing Service, 1964.
- KUDER, G. F., and M. W. RICHARDSON, The theory of the estimation of test reliability. *Psychometrika*, 1937, **2**, 151-160.
- LYERLY, S. B., The Kuder-Richardson formula (21) as a split-half coefficient and some remarks on its basic assumption. *Psychometrika*, 1958, **23**, 267-270.
- MEREDITH, W., Some results based on a general stochastic model for mental tests. *Psychometrika*, 1965, **30**, 419-440.
- NOVICK, M. R., and C. LEWIS, Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 1967, **32**, 1-13.
- RULON, P. J., A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 1939, **9**, 99-103.

## CHAPTER 5

# BASIC EQUATIONS OF THE CLASSICAL MODEL FOR HOMOGENEOUS TESTS OF VARIABLE LENGTH

### 5.1 Test Length as a Test Parameter

In Chapter 4, we derived a formula for the reliability of the composite measurement defined as the sum of two parallel measurements. We found that unless the component parts are completely unreliable, the composite measurement is more reliable than either of its component parts.

We may view this kind of result more broadly by imagining that the parallel measurements  $X_i$ ,  $i = 1, 2, \dots, n$ , are generated by  $n$  subtests and that the total test score of a randomly selected person is obtained by adding the subtest scores  $X_i$ . Thus if we define

$$X(n) = \sum_{i=1}^n X_i,$$

it is meaningful to think of the measurement  $X(n)$  as the score that a person obtains on a test (measurement) of *length*  $n$ . We shall therefore treat the length  $n$  of a test as a parameter of the test and study properties of the test in terms of variations of this parameter. In Section 4.5 a homogeneous test was defined as one having essentially  $\tau$ -equivalent components. In this chapter we study homogeneous tests whose components are, in fact, parallel.

More specifically, in this chapter we study the effect of test length on the reliability, validity, and other characteristics of a homogeneous test. Many of the results we shall derive or state in Sections 5.3 through 5.8 have been previously derived, and therefore our presentation in these sections will be somewhat brief. However, we shall work through a number of proofs so that the reader can become familiar with the manipulations employed in this new model, which has a notational generality that the classical model lacks but which is equivalent in its assumptions.

The main purpose of this chapter is to develop a very general approach to the study of test parameters as a function of test length. In contrast with the more classical formulation, this approach permits test length to be defined with equal ease either in integer units, as the number of items in the test, or as a continuous parameter, for example, the amount of time allowed for testing. This facilitates the treatment of speeded tests, as illustrated in Section 6.3 and in Chapter 21. It seems reasonable to speculate that, if large-scale testing programs should be-

come computerized, the concept of test length will need to be a more elastic one than that stated in Chapter 4.

This new statement of the classical model, which identifies the measurement process with the realization of a stochastic process, is due to Woodbury (1963). A more complete, self-contained treatment of this model has been given by Novick (1966). Some readers may be interested in noting that this model is, perhaps, the simplest case of a weakly stationary (covariance-stationary) stochastic process (Parzen, 1962). In restating Woodbury's model here, we have modified it in some technical respects to simplify it and to make it conform to the approach taken in this book; however, no essential features have been changed.

We shall discuss two very important results that have been obtained using this model. In Section 5.13 we give the most reliable composite with a specified true score, a composite which Woodbury and Lord derived with the help of this model. In Section 13.6 we give the relative time allocation that maximizes the multiple correlation. The algorithm for this optimum allocation was obtained by Woodbury and Novick, again using this model. We shall present a preview of this work in Section 5.16, in which the optimal allocation solution is given for a six-variable problem.

## 5.2 The Classical Model with a Continuous Test Length Parameter

If the items of a test yield parallel measurements and the test score is given by the sum of the discrete item scores, then we may write the following function of the test length:

$$X(n) = \sum_{i=1}^n X_i = \sum_{i=1}^n (T + E_i) = nT + \sum_{i=1}^n E_i, \quad n = 0, 1, 2, \dots \quad (5.2.1)$$

Suppose, however, that we associate the length of the test with a *continuous* parameter  $t > 0$ , which we call a *time parameter*. Then the equation analogous to (5.2.1) is

$$X(t) = tT + E(t), \quad (5.2.2)$$

where  $X(t)$  and  $E(t)$  are the observed- and error-score random variables at time (i.e., length)  $t \geq 0$ , and  $T$  is the true-score random variable at unit length. Indeed (5.2.2) can also cover the case of discrete  $t$ , if we simply define

$$E(t) = \sum_{i=1}^t E_i$$

and limit  $t$  to positive integral values. Furthermore, our manipulations on (5.2.2) will be done in a manner that will yield results which are mathematically valid for either the discrete or the continuous case.

Let us give an example of the flexibility of this model. Suppose we count the number of words a person reads correctly in  $t$  minutes and denote this by  $x(t)$ .

Then  $X(t)$  is the observed-score random variable corresponding to the scores obtained over repeated testings by randomly selected subjects. The expected value of the random variable for a fixed person is  $t\tau$ , and the true-score random variable  $tT$ , defined over persons, corresponds to the values  $t\tau$ . The error random variable is then  $E(t) = X(t) - tT$ .

On the other hand, we might obtain measurements by stopping each person after he had read  $t$  words correctly. Then the observed-score random variable  $X(t)$  would be the amount of time taken to accomplish these  $t$  word-readings. Although (5.2.2) is not usually interpreted for nonintegral  $t$  when applied to scores of ordinary paper-and-pencil tests, there are other kinds of measurements which might indeed depend on a continuous time parameter. For example,  $X(t)$  might be the number of tasks completed in a fixed time  $t$  (see Chapter 21). It is because of this greater scope that we are interested in the stochastic process model.

The mental testing applications of the results presented in this chapter are usually made at the subtest level and not at the item level, since approximately parallel items are difficult to obtain. For example, given a test with (approximately) known reliability, we might ask about the reliability of a test of double length ( $t = 2$ ) made up of the original test plus a parallel form. Also, assuming that the original test could be divided in half to produce two parallel subtest forms, we might wish to know what the reliability of a test of half length ( $t = \frac{1}{2}$ ) might be.

### 5.3 Statement of the Assumptions

We now restate the classical definitions and assumptions in a convenient form applicable to the continuous-time parameter model. Let

$$X(t) = tT + E(t), \quad (5.3.1)$$

where for fixed  $t$ , we have  $x(t)$ ,  $\tau$ , and  $e(t)$  as the values taken by the real-valued random variables  $X(t)$ ,  $T$ , and  $E(t)$  defined over a population of persons. The time parameter  $t$  may be defined either on the open interval  $(0, \infty)$  or on the positive integers; the statement of the definitions and assumptions and all derivations will apply to both cases. In this more general formulation, the assumptions of the classical model may conveniently be written as follows: When the intervals  $(t_1, t_2]$  and  $(t_3, t_4]$  are *disjoint*, that is, do not overlap, then

$$\mathcal{E}[E(t) | \tau] = 0, \quad \text{for all } \tau \text{ and all } t > 0, \quad (5.3.2)$$

$$\mathcal{E}[E(t_2) - E(t_1)]^2 = a^2 |t_2 - t_1|, \quad 0 < a < \infty, \quad (5.3.3)$$

$$\mathcal{E}\{[E(t_2) - E(t_1)][E(t_4) - E(t_3)]\} = 0, \quad (5.3.4)$$

where  $t_1$ ,  $t_2$ ,  $t_3$ , and  $t_4$  are values of the argument  $t$  and  $\mathcal{E}$  denotes expectation over persons. In this chapter, all expectations are defined over persons. The constant  $a$  is simply a scale factor determined by the scale of  $X(t)$ .

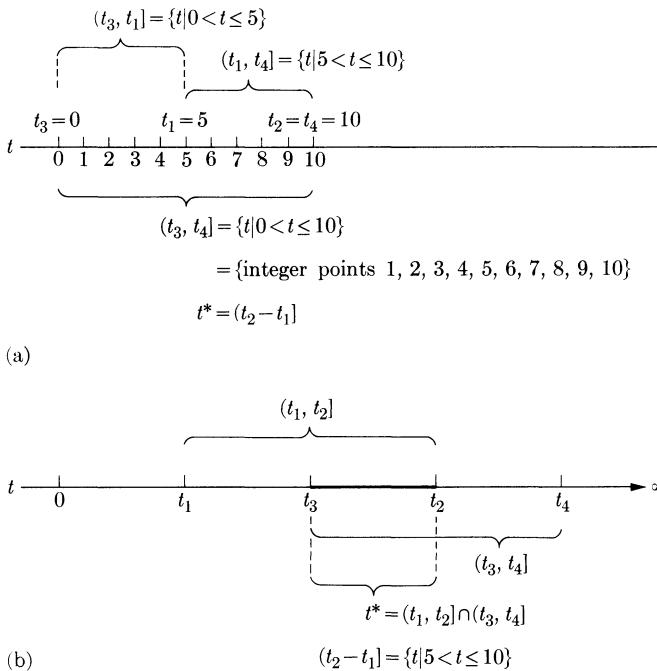


FIG. 5.3.1. Illustration of time intervals over which measurements may be defined for the (a) continuous and (b) discrete cases.

It is of some importance to note that (5.3.4) is stated in terms of left-open, right-closed intervals  $(t_1, t_2] = \{t \mid t_1 < t \leq t_2\}$ . In words, this notation means that the interval  $(t_1, t_2]$  is the set of all points  $t$  such that  $t$  is greater than  $t_1$  and  $t$  is less than or equal to  $t_2$ . Thus the right-hand endpoint  $t_2$  is part of the set but the left-hand endpoint  $t_1$  is not. In the discrete case,  $x(10) - x(5)$  is the measurement obtained in the interval  $(5, 10]$ ; since the point 5 is not part of the set,  $x(10) - x(5)$  is just the sum of the scores on items 6 through 10. See Fig. 5.3.1(a), where  $x(10) = x(10) - x(0)$  and  $x(5) = x(5) - x(0)$ , and the corresponding intervals are denoted by  $(t_3, t_4]$  and  $(t_3, t_1]$ .

From (5.3.3), we have  $\mathcal{E}[E(t_2) - 0]^2 \equiv \mathcal{E}[E(t_2)]^2 = a^2 t_2$ . But since  $\mathcal{E}[E(t_2)] = 0$ , we have  $\sigma^2[E(t_2)] = a^2 t_2$ . Then clearly  $\sigma^2[E(t)] \rightarrow 0$  as  $t \rightarrow 0$ . Hence  $E(t) \rightarrow 0$  in probability (or with probability 1) as  $t \rightarrow 0$ , and the definition  $E(0) = 0$  is therefore appropriate. By (5.3.1), this definition implies that  $X(0) = 0$ .

Assumption (5.3.2) states that the errors are unbiased. It corresponds to (2.7.4), which was originally obtained in the derivation of (2.7.1a). Hence (5.3.2) corresponds to (2.7.1a). It is a generalization of (2.7.1a), in that it states this property for all  $t > 0$ . An immediate implication of (5.3.2) is that  $\rho[E(t), T] = 0$  and  $\mathcal{E}[E(t)] = 0$ . Assumption (5.3.4) states that error scores on disjoint intervals are uncorrelated. It is a special case of (2.7.1c) in which the

disjoint  $t$  intervals are considered as distinct measurements. A generalized form of (5.3.4) covers measurements on distinct tests  $g$  and  $h$ ; for arbitrary subintervals  $(t_1, t_2]$  and  $(t_3, t_4]$ ,

$$\mathcal{E}\{[E_g(t_2) - E_g(t_1)][(E_h(t_4) - E_h(t_3))]\} = 0. \quad (5.3.4a)$$

Assumption (5.3.3) asserts that measurements defined on time intervals of equal length have equal error variance. The assumption (5.3.1), together with (5.3.2) and (5.3.3), is equivalent to the assumption that observed-score random variables defined on equal disjoint intervals are generated by parallel measurements. Assumption (5.3.4a) asserts that error scores at any length on distinct measurements are uncorrelated.

Because of the assumptions of the model, the properties (means, variances, covariances, etc.) of difference random variables defined on disjoint  $t$  intervals of the same length and pertaining to the same process are identical with one another and with those of such difference random variables generated by distinct but parallel processes. Suppose the intersection (common part) of the intervals  $(t_1, t_2]$  and  $(t_3, t_4]$ , denoted by

$$t^* = (t_1, t_2] \cap (t_3, t_4],$$

is null, that is, suppose these intervals are disjoint. Then let

$$\begin{aligned} X_g(t_g) &\equiv X(t_2) - X(t_1), & X_{g'}(t_{g'}) &\equiv X(t_4) - X(t_3), \\ t_g &\equiv t_2 - t_1, & t_{g'} &\equiv t_4 - t_3. \end{aligned}$$

This notation is justified since the difference  $X(t_2) - X(t_1)$  itself has the properties of a measurement. Then  $X_g(t_g)/t_g$  may be denoted simply as  $X_g/t_g$ . Thus, when measurements of the same length are defined on disjoint intervals, the use of the notation  $X_g$ ,  $X_{g'}$ , etc. is appropriate; such use will be made only under these conditions. Figure 5.3.1(b) illustrates an interval

$$t^* \equiv (t_1, t_2] \cap (t_3, t_4].$$

In Chapter 2 we introduced the concept of distinct measurements as experimentally independent measurements. We may now explicate this concept in the context of our present generalized version of the classical model. Consider the measurements  $X(t_4) - X(t_3)$  and  $X(t_2) - X(t_1)$ . If these overlap, say, if  $t_4 > t_2 > t_3 > t_1$ , then the measurement  $X(t_2) - X(t_3)$  is a part of each of the original measurements. As suggested by Theorem 5.5.1 and the basic statement  $X(t) = tT + E(t)$ , the relative size of  $t^* = t_2 - t_3$  determines the relative identity of the two measurements. In the extreme case in which  $t_4 = t_2$  and  $t_3 = t_1$ , we have

$$X(t_4) - X(t_3) \equiv X(t_2) - X(t_1).$$

For this model, if the time intervals of two measurements on the same process overlap, then the measurements are not distinct. If the endpoints of the measurements differ, then we say that the measurements are *partially distinct*, even though they may have a component in common (a *nondistinct* component). In Fig. 5.3.1(a) the measurements  $X(t_2) - X(t_1)$  and  $X(t_4) - X(t_3)$  are partially distinct even though they have the common component  $X(t_2) - X(t_3)$ . If measurements have no components in common, then they are *distinct*. If measurements are taken on two disjoint intervals of equal length, then all possible components of one measurement are distinct from all possible components of the second measurement. These measurements also have identical true scores and equal error variances. Hence, measurements on such intervals are parallel. We state this formally as

**Theorem 5.3.1.** Measurements on disjoint intervals of equal length on the same process are parallel measurements.

#### 5.4 The True Score as the Observed Score of a Person on a Test of Infinite Length

The true score of a person (whose ability remains unchanged) has often been defined as the observed score that a person would attain on a very long test. For the present model, we may note that the true score, as we have defined it, satisfies this condition. We note that  $\mathcal{E}[X(t)/t] = \tau$  and  $\sigma^2[X(t)/t] = a^2/t \rightarrow 0$  as  $t \rightarrow \infty$ . Hence the probability that  $X(t)/t$  differs from  $\tau$  by any amount (however small) is arbitrarily small for  $t$  taken sufficiently large. Hence  $X(t)/t$  converges to  $\tau$  in probability; i.e., the limiting value of  $X(t)/t$  as  $t \rightarrow \infty$  is  $\tau$ . In common language, then, we may *interpret* the true score as the (relative) observed score on a test of infinite length. This result is a special case of the weak law of large numbers. The advantage of the definition of true score adopted in this book is essentially one of mathematical and conceptual simplicity. Since from this definition and the assumptions of the model we have demonstrated the required asymptotic property, we lose nothing (semantically) by our definition.

#### 5.5 The Fundamental Theorem

One further simplification of the Woodbury model is possible. We present this simplification in the form of a theorem that we feel is of some intuitive value in acquiring an understanding of the nature of the model. We develop the proof of this theorem with some care, since all the manipulative techniques used in the algebra of this model are exhibited in this proof.

**Theorem 5.5.1.** Under the model specified above,

$$\mathcal{E}\{[E(t_2) - E(t_1)][E(t_4) - E(t_3)]\} = a^2 t^*, \quad (5.5.1)$$

where  $t^*$  is the interval length in the continuous case, or number of points in the discrete case, of  $(t_1, t_2] \cap (t_3, t_4]$ .

*Proof.* If  $t_4 = t_2$  and  $t_3 = t_1$ , then (5.5.1) reduces to (5.3.3); if  $(t_3, t_4] \cap (t_1, t_2]$  is the null set, then (5.5.1) reduces to (5.3.4). The proof for one special case in which the intervals overlap is typical of those for other cases. Suppose  $t_1 < t_3 < t_2 < t_4$ ; then, forming disjoint subintervals by adding

$$E(t_3) - E(t_1) \quad \text{and} \quad E(t_2) - E(t_3)$$

as necessary, we have

$$\begin{aligned} & \mathcal{E}\{[E(t_2) - E(t_1)][E(t_4) - E(t_3)]\} \\ &= \mathcal{E}\{[E(t_2) - E(t_3)] + [E(t_3) - E(t_1)]\} \{[E(t_4) - E(t_2)] + [E(t_2) - E(t_3)]\} \\ &= \mathcal{E}\{[E(t_2) - E(t_3)][E(t_4) - E(t_2)] + [E(t_3) - E(t_1)][E(t_4) - E(t_2)] \\ &\quad + [E(t_2) - E(t_3)][E(t_2) - E(t_3)] + [E(t_3) - E(t_1)][E(t_2) - E(t_3)]\} \\ &= a^2 t^*. \end{aligned}$$

The final expression is obtained by noting that (5.3.4) implies that the first, second, and fourth terms are zero and that (5.3.3) implies that the third term is equal to  $a^2 t^*$ .  $\square$

Thus the covariance of two error random variables of unit variance has a quite literal interpretation as the amount of variation the variables have in common, i.e., as the length of the interval over which the two errors overlap. The general method of proof is to divide the time interval into segments on which the measurements either coincide or do not overlap. To evaluate the expected value for the first type of interval we use (5.3.3). For the second type of interval we use (5.3.4).

## 5.6 Expectations and Variances

In the following theorem we develop the basic formulas for expectations and variances and define some test theory terms.

**Theorem 5.6.1.** Taking  $t_2 > t_1$ , we have

$$\mathcal{E}\left[\frac{X(t_2) - X(t_1)}{t_2 - t_1}\right] = \mathcal{E}\left(\frac{X_g}{t_g}\right) = \mu_T, \quad (5.6.1)$$

where  $\mu_T = \mathcal{E}T$ . We define two identities:

$$\sigma^2\left[\frac{E(t_2) - E(t_1)}{t_2 - t_1}\right] \equiv \sigma^2\left(\frac{E_g}{t_g}\right) = a^2(t_2 - t_1)^{-1}, \quad (5.6.2)$$

$$\sigma^2\left[\frac{X(t_2) - X(t_1)}{t_2 - t_1}\right] \equiv \sigma^2\left(\frac{X_g}{t_g}\right) = \sigma_T^2 + \sigma^2\left(\frac{E_g}{t_g}\right). \quad (5.6.3)$$

Taking  $t_2 = t$  and  $t_1 = 0$ , we have

$$\sigma^2[E(t)] = a^2t \quad (\text{the error variance}), \quad (5.6.4)$$

$$\sigma^2[E(t)/t] = a^2t^{-1} \quad (\text{the relative error variance}), \quad (5.6.5)$$

$$\sigma^2[X(t)] = t^2\sigma_T^2 + \sigma^2[E(t)] \quad (\text{the observed score variance}), \quad (5.6.6)$$

$$\sigma^2[X(t)/t] = \sigma_T^2 + \sigma^2[E(t)/t] \quad (\text{the relative observed score variance}). \quad (5.6.7)$$

*Proof.* We obtain these results by direct evaluation, using the techniques of Theorem 5.5.1 and noting that

$$\sigma^2[E(t)] = \sigma^2[E(t) - 0]. \quad \square$$

## 5.7 Covariances

Certain useful covariance formulas are given in

**Theorem 5.7.1.** For  $t_4 > t_3$  and  $t_2 > t_1$ , we have

$$\sigma\left[\frac{E(t_2)}{t_2}, \frac{E(t_1)}{t_1}\right] = a^2t_2^{-1}, \quad (5.7.1)$$

$$\sigma\left[\frac{X(t)}{t}, T\right] = \sigma_T^2, \quad (5.7.2)$$

$$\sigma\left[\frac{X(t)}{t}, \frac{E(t)}{t}\right] = \sigma^2\left[\frac{E(t)}{t}\right], \quad (5.7.3)$$

and

$$\sigma\left[\left(\frac{X(t_2) - X(t_1)}{t_2 - t_1}\right), \left(\frac{X(t_4) - X(t_3)}{t_4 - t_3}\right)\right] = \sigma_T^2 + \frac{a^2t^*}{(t_2 - t_1)(t_4 - t_3)}. \quad (5.7.4)$$

The reader should have no difficulty providing proofs of these results, using the techniques of proof of Theorem 5.5.1. In each case the proof is based on expressing the given quantities in terms of equivalent quantities defined over disjoint or coinciding subintervals. The following is an example of this procedure.

$$\begin{aligned} \sigma\left[\frac{E(t_2)}{t_2}, \frac{E(t_1)}{t_1}\right] &= \sigma\left[\frac{E(t_2) - E(t_1) + E(t_1)}{t_2}, \frac{E(t_1)}{t_1}\right] \\ &= \sigma\left[\frac{E(t_2) - E(t_1)}{t_2}, \frac{E(t_1)}{t_1}\right] + \sigma\left[\frac{E(t_1)}{t_2}, \frac{E(t_1)}{t_1}\right] \\ &= 0 + (t_1t_2)^{-1}\sigma^2[E(t_1)] \\ &= a^2t_2^{-1}. \end{aligned}$$

## 5.8 Correlations among Observed, True, and Error Scores

**Theorem 5.8.1.** For  $t_4 > t_3$  and  $t_2 > t_1$ , we have

$$\rho^2 \left( \frac{X(t)}{t}, T \right) = \frac{\sigma_T^4}{\sigma^2[X(t)/t]\sigma_T^2} = \frac{\sigma_T^2}{\sigma_T^2 + (a^2/t)}, \quad (5.8.1a)$$

$$\rho^2 \left[ \frac{X(t_2) - X(t_1)}{t_2 - t_1}, \frac{X(t_4) - X(t_3)}{t_4 - t_3} \right] = \frac{\left[ \sigma_T^2 + \frac{a^2 t^*}{(t_2 - t_1)(t_4 - t_3)} \right]^2}{\left[ \sigma_T^2 + \frac{a^2}{(t_2 - t_1)} \right] \left[ \sigma_T^2 + \frac{a^2}{(t_4 - t_3)} \right]}. \quad (5.8.1b)$$

We leave the derivation of an appropriate formula for  $\rho^2[X(t)/t, E(t)/t]$  as an exercise for the reader.

The limiting values of these correlations are given in

**Theorem 5.8.2**

$$\lim_{t \rightarrow \infty} \rho^2 \left[ \frac{X(t)}{t}, T \right] = 1, \quad \lim_{t \rightarrow 0} \rho^2 \left[ \frac{X(t)}{t}, T \right] = 0. \quad (5.8.2a)$$

$$\lim_{t \rightarrow \infty} \rho^2 \left[ \frac{X(t)}{t}, \frac{E(t)}{t} \right] = 0, \quad \lim_{t \rightarrow 0} \rho^2 \left[ \frac{X(t)}{t}, \frac{E(t)}{t} \right] = 1. \quad (5.8.2b)$$

$$\lim_{\substack{t_g \rightarrow \infty \\ t_{g'} \rightarrow \infty}} \rho^2 \left( \frac{X_g}{t_g}, \frac{X'_{g'}}{t_{g'}} \right) = 1, \quad \lim_{\substack{t_g \rightarrow 0 \\ t_{g'} \rightarrow 0}} \rho^2 \left( \frac{X_g}{t_g}, \frac{X'_{g'}}{t_{g'}} \right) = 0. \quad (5.8.2c)$$

The first result may be proved by letting  $t \rightarrow \infty$  in the right-hand member of Eq. (5.8.1a). We leave the proofs of the remaining results as exercises for the reader.

In words these results indicate that

- 1) the reliability of a test of infinite length is unity;
- 2) the reliability of a test of zero length is zero;
- 3) the correlation between observed and error scores for a test of infinite length is zero;
- 4) the correlation between observed and error scores for a test of zero length is unity;
- 5) the correlation between parallel forms of infinite length is unity;
- 6) the correlation between (parallel) forms of zero length is zero.

## 5.9 Expectations, Variances, and Correlations of Lengthened Tests

In this section, we consider measurements of unit length, that is,  $t = 1$ , and denote  $X(t)/t$  by  $X$ . We denote a measurement of length  $kt = k$  by  $X(k)$ . Using this notation, we give some basic results needed for further developments.

**Theorem 5.9.1**

$$\mathcal{E} \frac{X(k)}{k} = \mathcal{E} \frac{[k\mathbf{T} + E(k)]}{k} = \mu_X, \quad k = 1, 2, 3, \dots, \quad (5.9.1a)$$

$$\sigma^2 \left[ \frac{X(k)}{k} \right] = \frac{\sigma_X^2}{k} [1 + (k - 1)\rho_{XX'}], \quad (5.9.1b)$$

$$\sigma \left[ \frac{X(k)}{k}, X(1) \right] = \frac{1}{k} \sigma_X^2 [1 + (k - 1)\rho_{XX'}], \quad (5.9.1c)$$

where  $\rho_{XX'} = \rho[X(1), X'(1)]$ ,  $X(1)$  is a component of  $X(k)$ , and  $\sigma_X^2$  is the observed score variance at unit length.

*Proof.* Recalling that the measurements  $X_g$ ,  $g = 1, 2, \dots, k$ , are parallel, we have

$$\begin{aligned} & \sigma^2[X(k)/k] \\ &= \sigma^2 \{k^{-1}([X(1) - X(0)] + [X(2) - X(1)] + \dots + [X(k) - X(k - 1)])\} \\ &= \sigma^2 \left( k^{-1} \sum_{g=1}^k X_g \right) = k^{-2} \left[ \sum_{g=1}^k \sigma_X^2 + \sum_{g=1}^k \sum_{g'=1, g' \neq g}^k \sigma(X_g, X'_{g'}) \right] \\ &= k^{-2}[k\sigma_X^2 + k(k - 1)\sigma_X^2\rho_{XX'}] = \frac{\sigma_X^2}{k} [1 + (k - 1)\rho_{XX'}]; \\ & \sigma \left[ \frac{X(k)}{k}, X(1) \right] = \mathcal{E} \left[ \frac{X(k)}{k} \right] [X(1)] - \mathcal{E} \left[ \frac{X(k)}{k} \right] \mathcal{E}[X(1)] \\ &= k^{-1} \mathcal{E} \left( X_1^2 + X_1 \sum_{h=2}^k X_h \right) - \mu_X^2 \\ &= k^{-1} \left\{ (\mathcal{E}X_1^2 - \mu_X^2) + \left[ \mathcal{E} \left( X_1 \sum_{h=2}^k X_h \right) - (k - 1)\mu_X^2 \right] \right\} \\ &= k^{-1}\sigma_X^2[1 + (k - 1)\rho_{XX'}]. \quad \square \end{aligned}$$

**5.10 The Spearman-Brown Formula**

We now derive the general formula for the reliability of a lengthened test. Let  $X$  and  $X'$  be parallel measurements of unit length and  $X(k)$  and  $X'(l)$  be the corresponding measurements of lengths  $k$  and  $l$ , respectively. Then the *Spearman-Brown formula for the reliability of a lengthened test* can be written as

**Theorem 5.10.1**

$$\rho \left[ \frac{X(k)}{k}, \frac{X'(l)}{l} \right] = \frac{k\rho_{XX'}}{[1 + (k - 1)\rho_{XX'}]}. \quad (5.10.1)$$

*Proof*

$$\begin{aligned} \sigma \left[ \frac{X(k)}{k}, \frac{X'(l)}{l} \right] &= (kl)^{-1} \mathcal{E} \left[ \left( \sum_{g=1}^k X_g \right) \left( \sum_{h=1}^l X'_h \right) \right] - \mu^2 \\ &= (kl)^{-1}(kl\mathcal{E}XX') - \mu^2 = \sigma_X^2\rho_{XX'}, \end{aligned}$$

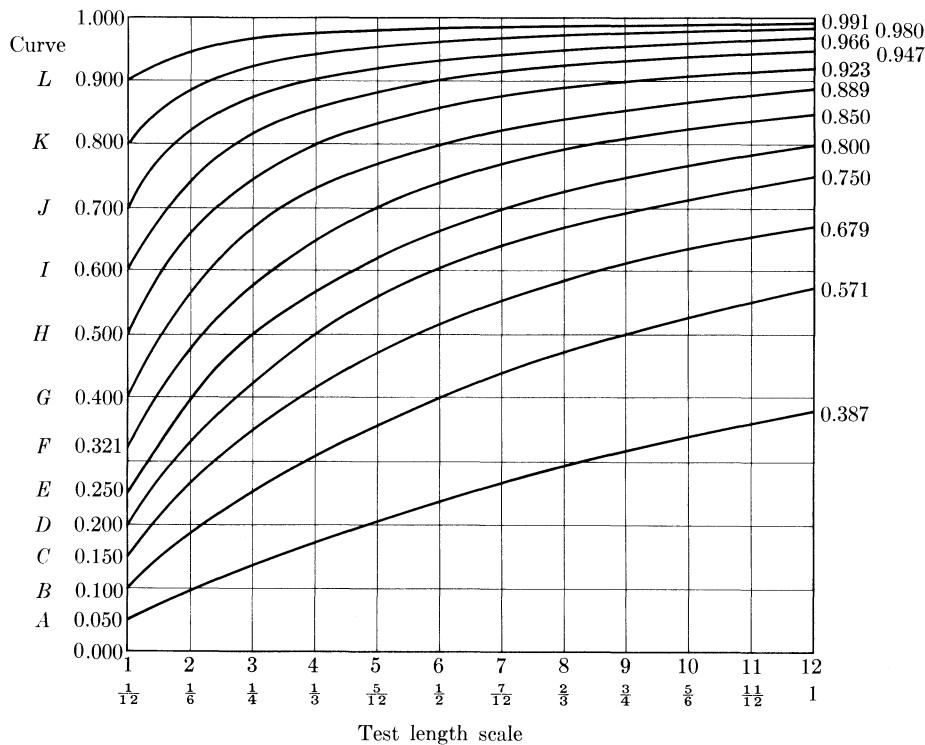


FIG. 5.10.1. Reliability as a function of test length.

**Table 5.10.1**  
Reliability as a function of test length

Test length scale

Curve	1	2	3	4	5	6	7	8	9	10	11	12
	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{5}{12}$	$\frac{1}{2}$	$\frac{7}{12}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{5}{6}$	$\frac{11}{12}$	1
L	.900	.947	.964	.973	.978	.982	.984	.986	.988	.989	.990	.991
K	.800	.889	.923	.941	.952	.960	.966	.970	.973	.976	.978	.980
J	.700	.824	.875	.903	.921	.933	.942	.949	.955	.959	.962	.966
I	.600	.750	.818	.857	.882	.900	.913	.923	.931	.938	.943	.947
H	.500	.667	.750	.800	.833	.857	.875	.889	.900	.909	.917	.923
G	.400	.571	.667	.727	.769	.800	.824	.842	.857	.870	.880	.889
F	.321	.486	.586	.654	.702	.739	.768	.791	.810	.825	.839	.850
E	.250	.400	.500	.571	.625	.667	.700	.727	.750	.769	.786	.800
D	.200	.333	.429	.500	.556	.600	.636	.667	.692	.714	.733	.750
C	.150	.261	.346	.414	.469	.514	.553	.585	.614	.638	.660	.679
B	.100	.182	.250	.308	.357	.400	.438	.471	.500	.526	.550	.571
A	.050	.095	.136	.174	.208	.240	.269	.296	.321	.345	.367	.387

which does not depend on  $k$  or  $l$ . This, together with (5.9.1b), establishes the theorem.  $\square$

In the proof of Theorem 5.10.1 it is assumed that  $k$  is a positive integer. If  $k$  is not integral but if, in lowest terms,  $k = a/b$ , where  $a$  and  $b$  are positive integers, then  $t/b$  may be taken as unit length and the above and all succeeding derivations are valid. If  $k$  is irrational, the results still hold by a limiting argument. Hence the theorem is valid for  $k > 0$ .

Equation (5.10.1) provides a method for determining what the reliability of a test will be after it is lengthened with parallel components, given only the reliability of the original test. If the components are  $\tau$ -equivalent but not parallel, then (4.4.4) should be used. Practical limitations on the usefulness of these formulas are discussed in Section 6.6.

Figure 5.10.1 illustrates the effect of test length on reliability. To determine the effect of lengthening a test on reliability, use the first line of the *test length scale*, on which the value at 1, the reliability at unit length, is given at the left. To determine the effect of shortening a test on reliability, use the second line, on which the same value is given at the right. Integral time values from 1 to 12 and fractional values in twelfths are given in Table 5.10.1. Intermediate values can be approximated from Fig. 5.10.1. For example, note that since curve  $B$  has a value of 0.400 at scale value 6, the values 0.250 and 0.182 of this curve at scale values 3 and 2 are the reliabilities of tests at  $\frac{1}{2}$  and  $\frac{1}{3}$  the length, respectively, of a test that has reliability of 0.400 at unit length. It is also possible to determine the length that a test must be to have a specified reliability. For example, curve  $E$  has a value of 0.400 at scale value 2, and a value of 0.800 at scale value 12; hence a test six times as long as the original test is required if the reliability is to be increased from 0.400 to 0.800. Techniques such as these may be used to increase the general usefulness of this table and figure.

### 5.11 The Effect of Test Length on Validity

More broadly, suppose that  $X$  and  $Y$  are generated by distinct processes. We then have

#### Theorem 5.11.1

$$\begin{aligned} \rho^2 \left[ \frac{X(k)}{k}, \frac{Y(l)}{l} \right] &= \frac{\sigma^2[X(k)/k, Y(l)/l]}{\sigma^2[X(k)/k]\sigma^2[Y(l)/l]} \\ &= \frac{k l \rho^2(X, Y)}{[1 + (k - 1)\rho(X, X')][1 + (l - 1)\rho(Y, Y')]} . \end{aligned} \quad (5.11.1)$$

This formula gives the validity of a lengthened (or shortened) test with respect to a second lengthened (or shortened) test. The function is strictly increasing in both  $k$  and  $l$ . The proof follows that of Theorem 5.10.1. Of all of the formulas of this chapter, this is the one most usefully committed to memory, because (5.10.1) and (5.11.2) are special cases of it and because of the results we shall obtain in the next section.

To see the effect of multiplying the length of one measurement by  $k$ , take  $l = 1$  in (5.11.1). The result gives the validity of a test of length  $k$  in terms of its validity at unit length. We state this result as

**Theorem 5.11.2**

$$\rho^2 \left[ \frac{X(k)}{k}, Y \right] = \frac{k\rho^2(X, Y)}{[1 + (k - 1)\rho(X, X')]} . \quad (5.11.2)$$

Figure 5.11.1 (a, b, c, d) illustrates the behavior of this curve for typical values of  $\rho(X, Y)$  and  $\rho(X, X')$ . The reader will wish to compare these curves with corresponding curves in Fig. 5.10.1. He should note that validity increases much more slowly with length than does reliability. He should also note that validity increases more quickly with length when the initial reliability is low, and decreases less quickly with length when the initial reliability is high.

By considering various limiting values as the time parameters become infinite in size, we are able to obtain the attenuation formulas of Section 3.9 as limiting cases of (5.11.1). The square of the correlation between two measurements when one is based on a test of infinite length (and hence is perfectly reliable) is

$$\begin{aligned} \rho^2(T_X, Y) &= \lim_{k \rightarrow \infty} \rho^2 \left[ \frac{X(k)}{k}, Y \right] \\ &= \lim_{k \rightarrow \infty} \frac{\rho^2(X, Y)}{1/k + [1 - (1/k)]\rho(X, X')} = \frac{\rho^2(X, Y)}{\rho(X, X')} . \end{aligned} \quad (5.11.3)$$

Similarly

$$\rho^2(X, T_Y) = \lim_{l \rightarrow \infty} \rho^2 \left[ X, \frac{Y(l)}{l} \right] = \frac{\rho^2(X, Y)}{\rho(Y, Y')} . \quad (5.11.4)$$

The square of the correlation between two measurements when both are of infinite length (and hence perfectly reliable) is

$$\rho^2(T_X, T_Y) = \lim_{\substack{k \rightarrow \infty \\ l \rightarrow \infty}} \rho^2 \left[ \frac{X(k)}{k}, \frac{Y(l)}{l} \right] = \frac{\rho^2(X, Y)}{\rho(X, X')\rho(Y, Y')} . \quad (5.11.5)$$

Formulas (5.11.3), (5.11.4), and (5.11.5) are called attenuation formulas, or corrections for attenuation. For example, we see from (5.11.1) and (5.11.5) that the correlation  $\rho[X(k)/k, Y(l)/l]$  increases with  $k$  and  $l$  and reaches its maximum when  $k = l = \infty$ , when its value is  $\rho(T_X, T_Y)$ , the correlation between the corresponding true scores. The correlation  $\rho(X, Y)$  is said to be *attenuated* by the unreliability of each measurement. The value  $\rho(T_X, T_Y)$  is called the *disattenuated correlation*; it is the correlation between the measurements when they are perfectly reliable. Equation (5.11.4) places an upper limit on the correlation which can be obtained between two measurements by reducing the unreliability of each by an increase of its length. Similarly (5.11.3) places an upper limit on the correlation which can be obtained when just one

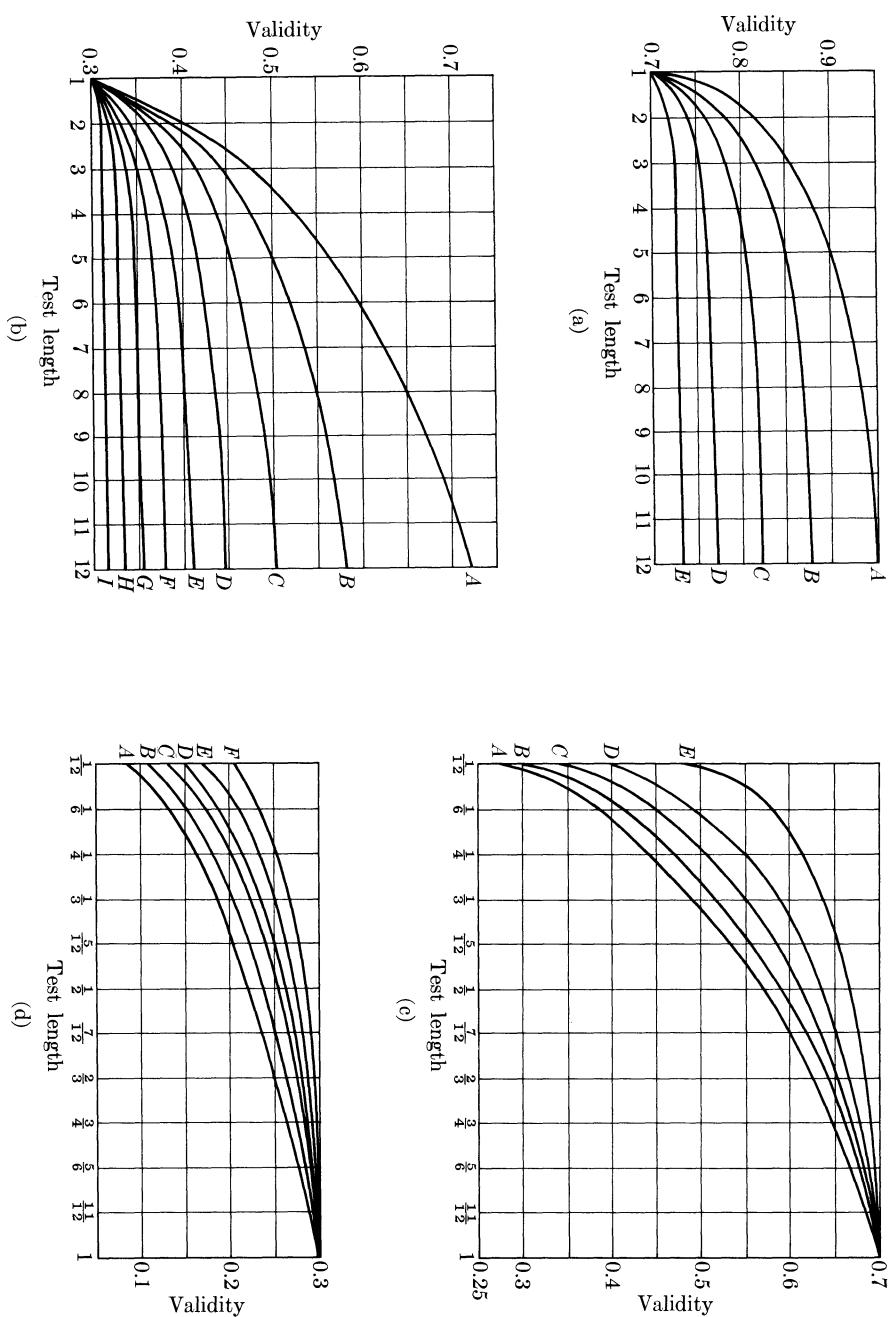


FIG. 5.11.1. Effect of test length on validity.

**Table for Part (c)**

Initial validity = 0.7			
Curve	Initial reliability $k = \frac{1}{12}$	Reliability at $k = \frac{1}{12}$	Validity at $k = \frac{1}{12}$
A	0.5	0.923	0.951
B	0.6	0.947	0.880
C	0.7	0.966	0.822
D	0.8	0.980	0.775
E	0.9	0.991	0.735

**Table for Part (b)**

Initial validity = 0.3

Initial validity = 0.3			
Curve	Initial reliability $k = 12$	Reliability at $k = 12$	Validity at $k = \infty$
A	0.1	0.571	0.717
B	0.2	0.750	0.581
C	0.3	0.837	0.501
D	0.4	0.889	0.447
E	0.5	0.923	0.408
F	0.6	0.947	0.377
G	0.7	0.966	0.352
H	0.8	0.980	0.332
I	0.9	0.991	0.315

**Table for Part (d)**

Initial validity = 0.3

Initial validity = 0.3			
Curve	Initial reliability $k = \frac{1}{12}$	Reliability at $k = \frac{1}{12}$	Validity at $k = \frac{1}{12}$
A	0.1	0.009	0.091
B	0.4	0.053	0.109
C	0.6	0.111	0.129
D	0.7	0.163	0.145
E	0.8	0.250	0.168
F	0.9	0.429	0.207

measurement is made more reliable. The extreme equalities in (5.11.3) through (5.11.5) were obtained in Section 3.9. Here, however, their character as upper bounds on validity is made clear. In practice, some care must be used in applying these formulas, as explained in Section 6.5.

### 5.12 Comparing Reliabilities and Validities of Tests of Differing Lengths

If one test takes two hours to administer and has a reliability of 0.80 and a second takes only one hour and has a reliability of 0.75, we might well wish to apply the Spearman-Brown formula to determine what the reliability of the second test would be if it were doubled in length and took two hours to administer. A simple computation indicates that the second test would have a reliability of 0.86 and hence would be more reliable than the first test if each had an equal length of two hours.

We may wish to determine the intrinsic reliabilities of a number of tests of different lengths, i.e., the relative reliabilities they would have, were they of equal length. One simple way of doing this is to use the Spearman-Brown formula in reverse to determine, for each test, what its reliability would be at unit length. We may easily obtain

$$\rho = \frac{\rho_{k'}}{k' - (k' - 1)\rho_{k'}}, \quad (5.12.1)$$

where  $\rho$  is the reliability of the test at unit length and  $\rho_{k'}$  is its reliability at length  $k'$ . Computing values of  $\rho$  for each test, we can determine which would have the greatest reliability at unit length. Since the Spearman-Brown formula determines the reliability of a test at length  $k$  as a strictly increasing function of its reliability at unit length and the value  $k$ , it is clear that the ordering of the reliabilities of the various tests must be the same at length  $k$  as at unit length. An equivalent approach to this problem is suggested in Exercise 5.3.

Suppose four verbal ability tests are available with test lengths (in numbers of items) and corresponding reliabilities (61, 0.85), (121, 0.90), (141, 0.95), (211, 0.96). We might desire to know which test type would be most reliable if the tests were of equal length. Applying (5.12.1), we determine that the reliability of the first test at unit length is

$$\rho = \frac{0.85}{61 - 60(0.85)} = 0.085.$$

Similarly we determine the reliabilities of the other tests of unit length to be 0.069, 0.119, and 0.102, respectively. Thus *for any fixed length* the third test will be most reliable.

From (5.10.1) it is apparent that the reliability of a test is not directly proportional to its length. However, there is a monotonic function of  $\rho_{XX'}$  which is proportional to the length of the test, namely, the *signal-to-noise ratio*  $\rho/(1 - \rho)$ , where  $\rho = \rho_{XX'} = \rho_{XT}^2$ . Let  $\rho_k$  and  $\rho_{k'}$  be the reliabilities of the test at lengths  $k$  and  $k'$ . We then have the following theorem.

**Theorem 5.12.1**

$$\frac{\rho_k/(1 - \rho_k)}{\rho_{k'}/(1 - \rho_{k'})} = \frac{k}{k'}. \quad (5.12.2)$$

*Proof.* From Exercise 5.2 we have

$$\rho_k = \frac{k\rho_{k'}}{k' + (k - k')\rho_{k'}}, \quad (5.12.3)$$

which is the reliability of a test at length  $k$  in terms of the reliability at length  $k'$ . The required result is then obtained by simple algebra.  $\square$  Equation (5.12.1) can be obtained from (5.12.3) by setting  $k = 1$ .

The signal-to-noise ratio may be rewritten as

$$\frac{\rho}{1 - \rho} = \frac{\sigma_T^2/\sigma_X^2}{1 - (\sigma_T^2/\sigma_X^2)} = \frac{\sigma_T^2}{\sigma_E^2},$$

that is, as the ratio of the true-score variance (the signal power) to the error-score variance (the noise power). Cronbach and Gleser (1964) have discussed the analogy of reliability theory to communication theory. They have pointed out the relation between the signal-to-noise ratio and the length of a test, and have given table values of the signal-to-noise ratio corresponding to values of  $\rho$  and indicated that this table may be used as a substitute for the Spearman-Brown formula. We may solve (5.12.2) for  $k$  and obtain

**Theorem 5.12.2.** If a test at length  $k'$  has reliability  $\rho_{k'}$ , the length  $k$  that it must have in order to have a specified reliability  $\rho_k$  is

$$k = k' \frac{\rho_k/(1 - \rho_k)}{\rho_{k'}/(1 - \rho_{k'})}. \quad (5.12.4)$$

Suppose that the reliability of a 90-minute test is 0.80 and that we desire the test to have a reliability of 0.90. Substituting in (5.12.4), we obtain

$$k = 90 \frac{0.90/(1 - 0.90)}{0.80/(1 - 0.80)} = 202.5$$

as the required length of the test in minutes. A comparable treatment of validity is given in the exercises for this chapter.

**5.13 The Most Reliable Composite with a Specified True Score\***

An important problem for which the stochastic process formulation provides a convenient framework is that of determining the most reliable composite with a specified true score. If a composite score is to be obtained from a battery of

---

\* Reading of this section may be omitted without loss of continuity; however, no advanced mathematical techniques are employed.

tests, the manner in which these tests enter into the composite may be varied in two respects: (1) Different weights may be assigned to the tests, and (2) the test lengths may be altered. Weights and administration times may be assigned so as to maximize either the validity or the reliability of the composite. We shall discuss the first of these alternatives in Chapter 13. In the present section we present a simple method, due to Woodbury and Lord (1956), for maximizing reliability. At the end of this section we shall give the solution to the simple problem of maximizing reliability when the test lengths but not the test weights are considered fixed.

If we place no restrictions on the composite, then the solution to the problem of this section becomes trivial, since in the optimal composite all the available time will be given to the test in the battery having the greatest reliability at unit (and hence at any other specified) length. We can obtain an optimal solution of practical value by specifying the true score on the composite and holding this fixed while the reliability is maximized.

We consider the weighted composite score

$$Y(t) = \sum_{i=1}^n w_i x_i(t_i). \quad (5.13.1)$$

The true score  $\eta$  of this composite at unit length is proportional to the corresponding composite of true scores:

$$\eta \propto \sum_i w_i t_i \tau_i. \quad (5.13.2)$$

The problem is to maximize the reliability  $\rho_{YY'}$ , subject to the following restrictions:

1. The total administration time  $t = \sum_{i=1}^n t_i$  is limited to a maximum value  $T > 0$ :

$$t = \sum_{i=1}^n t_i \leq T. \quad (5.13.3)$$

2. The testing time for each component is nonnegative:

$$t_i \geq 0, \quad i = 1, 2, \dots, n. \quad (5.13.4)$$

3. The true score of the composite is specified except for an unknown coefficient of proportionality  $p$ :

$$\eta \propto \sum_i w_i t_i \tau_i = p \sum_i \gamma_i \tau_i, \quad (5.13.5)$$

where the  $\gamma_i = w_i t_i / p$  are specified constants and  $p$  is a fixed coefficient of proportionality.

We further assume that none of the component measurements are essentially  $\tau$ -equivalent.

The relative values of  $\gamma_i$  must be specified in advance on the basis of subjective judgment or special information. One method of doing this is to estimate the administration times that would yield the desired composite true score for an unweighted composite. If this can be done, then the administration times chosen may be treated as the required values of  $\gamma_i$ . In practice, a frequent application may be to determine how the composite score on a given battery can be made more reliable without changing its true score.

The solution to this problem of maximizing reliability is given as

**Theorem 5.13.1.** Consider  $n$  tests with observed scores  $X_i(t_i)$ , lengths  $t_i$ , reliabilities  $\rho_{ii'}(t_i)$ , and variance-covariance matrix  $\|\sigma[X_i(t_i), X_j(t_j)]\|$  at length  $t_i$ ,  $i = 1, 2, \dots, n$ . Assume a maximum total testing time  $T$ , and  $n$  constants  $\gamma_i$  defining the true score of the desired composite. Then the assignment of test lengths and test weights that maximizes the reliability of the composite for the specified true score can be determined by the following procedure:

1. For each test, compute the standard error of measurement at unit length from the formula

$$\sigma(E_i) = t^{-1/2} \sigma[X_i(t_i)][1 - \rho_{ii'}(t_i)]. \quad (5.13.6)$$

2. Compute the new testing time for each test from the formula

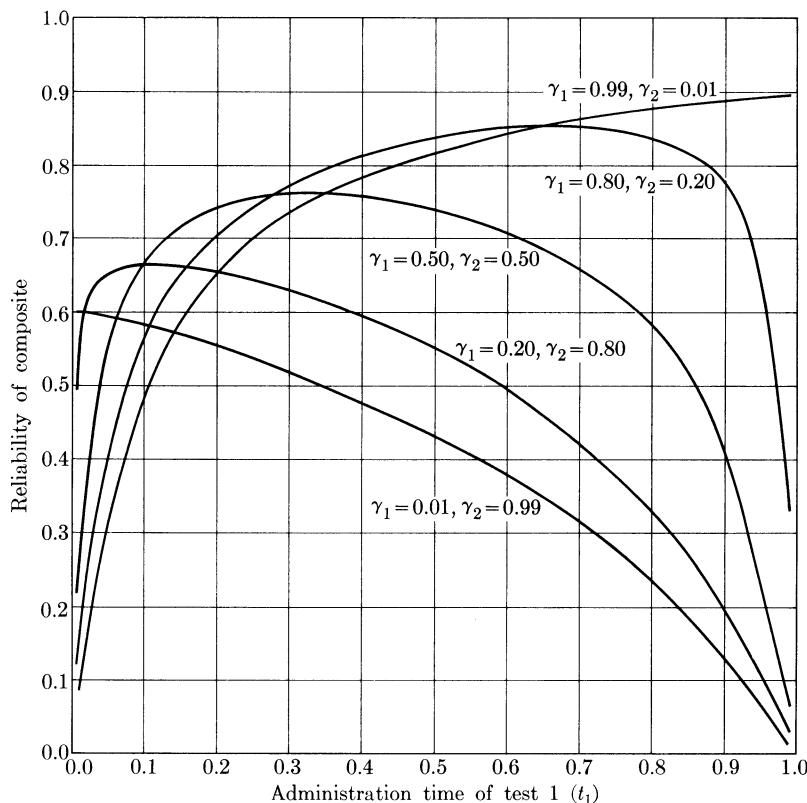
$$\tilde{t}_i = \frac{T}{\sum_i |\gamma_i| \sigma(E_i)} |\gamma_i| \sigma(E_i). \quad (5.13.7)$$

3. Compute the new test weight from the formula

$$\tilde{w}_i = \frac{\gamma_i}{|\gamma_i|} [\sigma(E_i)]^{-1}. \quad (5.13.8)$$

4. Compute the reliability of the composite from formulas (13.5.1d) and (4.7.2).

From (5.13.7) we see that the available testing time assigned to each test in the optimal battery is a function of the standard error of measurement of that test at unit length *and* of the desired composite true score. This proportion *does not* depend on total testing time available, nor on any of the parameters of the other tests in the battery, except insofar as these parameters influence the specified composite true score. From (5.13.8) we see that, except for sign, the weight assigned to any test in the optimal composite is determined *solely* by the standard error of measurement of that test at unit length. If we scale each test to have the same  $\sigma_E$  at unit length, then *any* unweighted composite has maximum reliability for measuring its own true score. From (5.13.8) we see also that if all tests have equal standard errors of measurement at unit length, then the *absolute values* of the weights and test lengths cannot be altered without either lowering the reliability of the composite or changing its true



Reliability of an unweighted composite and of the optimally weighted composite having the same true score

True Score	Unweighted composite			Optimally weighted composite				
	$\gamma_1$	$\gamma_2$	$t_1$	$w_1$	$r_{yy}(1)$	$\hat{t}_1$	$\hat{w}_1$	$r_{yy}(1)$
.20 .80	.20	1	.654			.111	3.2	.665
.50 .50	.50	1	.744			.333	3.2	.763
.80 .20	.80	1	.837			.667	3.2	.851

FIG. 5.13.1. The reliability of all possible composites for each of 5 different true scores, shown as a function of the allocation of testing time. [From Woodbury and Lord, The most reliable composite with a specified true score. *British Journal of Statistical Psychology*, 1956. Used by permission.]

score. Incidentally it may be shown that the optimal composite has the greatest advantage over the usual unweighted composite in the case in which the individual tests have greatly varying standard errors of measurement at unit length.

We omit the proof of the theorem of this section since the theorem is of a very highly specialized nature and since its proof is lengthy and involves rather specialized techniques. It may be found in Woodbury and Lord (1956). Instead, and more profitably, we present a figure and table from that source that show the reliability of a two-component composite as a function of relative test length for various prespecified true scores.

Figure 5.13.1 illustrates how composite score reliability varies when the two components have unit observed-score variance, reliabilities  $\rho_{11'} = 0.90$  and  $\rho_{22'} = 0.60$ , and intercorrelation  $\rho_{12} = 0.70$ , all at unit length. For each of five different true scores, as specified by values of  $\gamma_1$  and  $\gamma_2$ , the figure shows the reliability  $\rho_{YY'}$  of the composite at unit length as a function of the proportion ( $t_1$ ) of total testing time ( $t_0 = 1$ ) allocated to test 1.

The figure also shows that when the true composite score involves the component true scores in the ratio 4:1 ( $\gamma_1 = 0.80$ ,  $\gamma_2 = 0.20$ ), the maximum reliability is attained when  $t_1 = 0.667$ . The unweighted composite having the same true score is obtained when  $t_1 = 0.8$  and  $t_2 = 0.2$ . The reliability of this unweighted composite is found to be 0.837. This value is numerically only slightly less than that of the optimally weighted composite having the same true score ( $\rho_{YY'} = 0.851$ ); however, the length of the unweighted composite would have to be increased 11% in order to increase the reliability of the composite from 0.837 to 0.851. The reliabilities of other unweighted and optimally weighted composites are compared in the table of Fig. 5.13.1.

Figure 5.13.1 demonstrates that for the components considered and for each of the composite true scores considered, the reliability is heavily dependent on the relative test lengths. This heavy dependence results from the fact that the two composites have greatly different reliabilities at unit lengths (0.90 and 0.60) for the two components. Less variant reliabilities would produce much flatter curves.

## 5.14 Maximizing the Reliability of the Composite When Component Lengths Are Fixed\*

A simpler problem, treated by Thomson (1940), Mosier (1943), Peel (1948), and Green (1950), is that of determining weights so as to maximize the reliability  $\rho$  of the composite when the component lengths are considered fixed. Let  $\mathbf{R}$  be the intercorrelation matrix of the components, and  $\mathbf{R}^*$  a matrix whose off-diagonal elements are the same as those of  $\mathbf{R}$  but whose diagonal elements are the respective reliabilities of the components. Then the required weights can be obtained by using the following theorem.

---

\* Reading of this section may be omitted without loss of continuity. The proof given herein depends on the theory of latent roots and vectors of matrices.

**Theorem 5.14.1.** If  $\mathbf{R}$  and  $\mathbf{R}^*$  are defined as above, any vector  $\mathbf{u}$  of constants whose elements are proportional to the latent vector corresponding to the first latent root of the matrix  $\mathbf{R}^*\mathbf{R}^{-1}$  is a vector of optimal weights.

*Proof.* Let  $\mathbf{u}$  be a  $k$ -element solution vector. Then, by (4.7.2),

$$\rho = \frac{\mathbf{u}'\mathbf{R}^*\mathbf{u}}{\mathbf{u}'\mathbf{R}\mathbf{u}}. \quad (5.14.1)$$

Since the reliability of the composite remains unchanged if all weights are multiplied by a constant, we may impose the restriction that the variance of the composite be unity. Then the quantity to maximize is

$$\mathbf{u}'\mathbf{R}^*\mathbf{u} - \lambda(\mathbf{u}'\mathbf{R}\mathbf{u} - 1). \quad (5.14.2)$$

Differentiating with respect to  $\mathbf{u}'$  and setting the derivative equal to  $\mathbf{0}'$  yields

$$\mathbf{u}'(\mathbf{R}^* - \lambda\mathbf{R}) = \mathbf{0}'. \quad (5.14.3)$$

This system has a solution only if the determinant of the coefficients of  $\mathbf{u}'$  is zero:

$$|\mathbf{R}^* - \lambda\mathbf{R}| = 0. \quad (5.14.4)$$

Equation (5.14.4) defines the characteristic equation of the matrix  $\mathbf{R}^*\mathbf{R}^{-1}$ . If we solve this polynomial equation for its largest root and determine the corresponding vector  $\mathbf{u}'$  from (5.14.3), we obtain a solution vector. However, multiplication of this vector by any constant yields an equally valid solution.  $\square$

## 5.15 Maximizing the Validity of a Test Battery as a Function of Relative Test Lengths for a Fixed Total Testing Time

The most complex, and in many respects the most interesting, problem in classical test theory is that of adjusting relative test lengths to maximize the multiple correlation of a test battery with a fixed criterion. The problem is of great interest because the solution depends so sensitively on the intercorrelations of the variables, their validities, and their reliabilities. In Section 13.6 we shall sketch the solution algorithm. In this section we present the solution to a five-variable problem, using data gathered by French (1963). The variables are five high-level predictor variables based on item types not found in standard aptitude tests. The reliabilities of the variables at unit length are 0.76, 0.82, 0.58, 0.70, and 0.64. The validities at unit length are 0.44, 0.15, 0.39, 0.30, and 0.40. The intercorrelation matrix is

$$\begin{vmatrix} 0.13 & 0.31 & 0.28 & 0.32 \\ 0.31 & 0.15 & 0.09 & 0.15 \\ 0.28 & 0.09 & 0.19 & 0.32 \\ 0.32 & 0.15 & 0.32 & 0.33 \end{vmatrix}.$$

**Table 5.15.1**

General solution to the relative time allocation problem for the French data

Variable	$T \leq 0.20$	$0.20 \leq T \leq .42$	$0.42 \leq T \leq 1.45$	$1.45 \leq T \leq 8.50$	$T \geq 8.50$
1	$T$	$0.20 + .435(T - .20)$	$0.30 + .279(T - .42)$	$0.59 + .253(T - 1.45)$	$2.37 + .251(T - 8.50)$
2	0.00	0.00	0.00	0.00	$0.00 + .012(T - 8.50)$
3	0.00	0.00	$0.00 + .398(T - .42)$	$0.41 + .386(T - 1.45)$	$3.13 + .380(T - 8.50)$
4	0.00	0.00	0.00	$0.00 + .092(T - 1.45)$	$0.65 + .092(T - 8.50)$
5	0.00	$0.00 + .565(T - .20)$	$0.12 + .323(T - .42)$	$0.45 + .269(T - 1.45)$	$2.35 + .265(T - 8.50)$

The general solution to the time allocation problem for total testing times  $T$  is given in Table 5.15.1.

An interesting feature of the solution is that for small values of  $T$ , variable 1 receives the entire time allocation, while for large values of  $T$ , variable 3 receives the larger time allocation. The reason for this is that variable 1 is more reliable than variable 3 and hence suffers less from shortening than does variable 1. Exercise 5.9 suggests that if validities are equal for two tests of equal length, the more reliable one should be used if a shortened form is required and the less reliable one should be used if a lengthened form is required.

## Exercises

### 5.1.

Test	Mean observed score	Observed score standard deviation	Number of items	Reliability	Validity
A	16.5	4.4	30	0.72	0.68
B	12.6	3.5	20	0.77	0.50
C	66.3	17.2	120	0.95	0.75
D	52.2	13.7	100	0.95	0.77

This table gives statistics on four tests proposed as predictors of grade-point average. The samples on which these statistics are based were very large. The reliability of the criterion has been estimated (again, on the basis of a very large sample) to be 0.72. We presume the test length to be equal to the number of items.

- If test A is lengthened to a 100-item test, what will be the new mean observed score, observed-score standard deviation, reliability, and validity?
- If test B is to be lengthened to increase its reliability to 0.90, how many new items are needed? What will the new validity be?
- Which of the four tests, in its present form, is best for use in predicting the criterion?
- Assuming there is time for a 200-item test, which test would be most valid at that length?

- e) Which test has the highest validity, disattenuated for unreliability in both the predictor and the criterion?
- 5.2. Show that the reliability  $\rho_k$  of a test at length  $k$  in terms of its reliability  $\rho_{k'}$  at length  $k'$  is
- $$\rho_k = \frac{k\rho_{k'}}{k' + (k - k')\rho_{k'}}.$$
- 5.3. Demonstrate that  $k(1/\rho_k - 1) = k'(1/\rho_{k'} - 1)$  for arbitrary lengths  $k$  and  $k'$ , and hence that the quantity  $l(1/\rho_l - 1)$  is invariant with respect to test length (Gulliksen, 1950, p. 83). Discuss how this quantity is related to the signal/noise ratio and how it may be used to compare the relative reliabilities of tests of differing lengths.
- 5.4. By correcting the correlation  $\rho(X, X')$  between parallel measurements  $X$  and  $X'$  for attenuation, show that  $\rho_{XT}^2 = \rho_{XX'}$ .
- 5.5. Let  $\rho_k$  be the reliability of a test at length  $k$ , and let  $\Delta_k = \rho_k - \rho_1$  be the increase in reliability due to lengthening a test to  $k$  times its original length. Show that the maximum value of this increment is attained for

$$\rho_1 = \frac{-1 \pm \sqrt{k}}{k - 1}.$$

- 5.6. If the reliability of a test is increased from 0.80 to 0.90 by lengthening the test, what is the validity of the lengthened test if its original validity was 0.65?
- 5.7. Let  $\rho_{XY}$  be the validity of measurement  $X$  with respect to measurement  $Y$ , and  $\rho_{XX'}$  be the reliability of measurement  $X$ . Suppose that measurement  $X$  is lengthened and its new (increased) reliability is denoted by  $\rho'_{XX'}$ . The validity of  $X$  with respect to  $Y$  will also be increased. Denote this increased validity by  $\rho'_{XY}$ . (Thus we have  $\rho'_{XX'} > \rho_{XX'}$ ,  $\rho'_{XY} > \rho_{XY}$ ). Show that

$$\rho(X, Y)/\rho'(X, Y) = \sqrt{\rho(X, X')/\rho'(X, X')}.$$

- 5.8. From the result of the above exercise, show that

$$\frac{\rho[X(l), Y]}{\sqrt{\rho[X(l), X'(l)]}}$$

is invariant with respect to test length.

- 5.9. Let  $X_1$  and  $X_2$  be measurements with reliabilities  $\rho^2(X_1, T_1) > \rho^2(X_2, T_2)$  and equal validities  $\rho(X_1, Z) = \rho(X_2, Z)$ , with a specified criterion  $Z$  at unit length. Show that at length  $t$ ,

$$\begin{aligned} \rho[X_1(t), Z] &< \rho[X_2(t), Z] & \text{for } t > 1, \\ \rho[X_1(t), Z] &> \rho[X_2(t), Z] & \text{for } t < 1. \end{aligned}$$

State in words the principle to be inferred from this result.

- 5.10. Now consider the case in which the validities of the two tests are different. Suppose test A has a validity coefficient of 0.66 and a reliability of 0.72, and

test B has a validity of 0.70 and a reliability of 0.90. If each test has unit length, at what length will the tests have equal validity? What would the validities of test A and test B be if both tests were increased to infinite length?

- 5.11. Show that the validity  $\rho_k[X(k)/k, Y]$  of a test at length  $k$  may be written in terms of its validity  $\rho_{k'}[X(k')/k', Y]$  and reliability  $\rho_{k'}(X, X')$  at length  $k'$  as

$$\rho_k\left(\frac{X(k)}{k}, Y\right) = \rho_{k'}\left(\frac{X(k')}{k'}, Y\right) \sqrt{\frac{k}{k' + (k - k')\rho_{k'}(X, X')}}.$$

- 5.12. Show that the variance  $\sigma^2[X(k)/k]$  of a test at length  $k$  may be written in terms of its reliability  $\rho_{k'}(X, X')$  and variance  $\sigma^2[X(k')/k']$  at length  $k'$  as

$$\sigma^2\left(\frac{X(k)}{k}\right) = \sigma^2\left(\frac{X(k')}{k'}\right) \frac{k' + (k - k')\rho_{k'}(X, X')}{k}.$$

- 5.13. Let  $X_1$  and  $X_2$  be measurements with reliabilities  $\rho^2(X_1, T_1) > \rho^2(X_2, T_2)$  and validities  $\rho(X_1, Z)$  and  $\rho(X_2, Z)$ , with a specified criterion  $Z$  at unit length. If  $a\rho(X_1, Z) = \rho(X_2, Z)$ , where  $a < 1$ , show that the validities will be equal at length

$$t = 1 + \frac{1 - a^2}{a^2\rho^2(X_1, T_1) - \rho^2(X_2, T_2)}.$$

Also, show that at values of  $t$  greater than this value,  $\rho[X_1(t), Z] < \rho[X_2(t), Z]$ . Show that this inequality is reversed for  $t$  less than this value.

- 5.14. Using Theorem 5.9.1, show that the correlation of a homogeneous test of length  $l$  with a subtest of length  $n < l$  is given by

$$\rho^2[X(l), X(n)] = \frac{\rho + (1 - \rho)/l}{\rho + (1 - \rho)/n},$$

where  $\rho$  is the reliability of  $X$  at unit length.

- 5.15. If  $\rho_{kl}[X(k)/k, Y(l)/l]$  is the correlation of the relative observed scores  $X(k)/k$  and  $Y(l)/l$  at lengths  $k$  and  $l$ , and  $\rho_k[X(k)/k, X'(k)/k]$  and  $\rho_l[Y(l)/l, Y'(l)/l]$  are the reliabilities of  $X(k)/k$  and  $Y(l)/l$  at lengths  $k$  and  $l$ , show that the correlation between  $X(k')/k'$  and  $Y(l')/l'$  at lengths  $k'$  and  $l'$  is

$$\rho\left[\frac{X(k')}{k'}, \frac{Y(l')}{l'}\right] = \frac{\sqrt{kk'l'}}{[1 + (kk' - 1)\rho_{XX'}]^{1/2}[1 + (ll' - 1)\rho_{YY'}]^{1/2}} \cdot \rho_{XY}.$$

- 5.16. Let  $\rho_k$  be the reliability of a test at length  $k$ , and let  $\delta_k = \sqrt{\rho_k} - \sqrt{\rho_1}$  be the increase in the index of reliability due to lengthening a test to  $k$  times its original length. Find the value  $\rho_1$  for which  $\delta_k$  is a maximum. Comment on the relationship between this result and that of Exercise 5.5. Which result is more meaningful?

- 5.17. Show that in the context of Section 5.4,  $X(t)/t \rightarrow \tau$  with probability one.

- 5.18. In the context of Section 5.3, show that as  $t \rightarrow 0$ ,  $E(t)/t$  converges to a random variable having zero mean and infinite variance, while for fixed  $\tau$ , as  $t \rightarrow 0$ ,  $X(t)/t$  converges to a random variable having mean  $\tau$  and infinite variance.

- 5.19. Let  $\rho_{XY}^2$  be the squared validity of  $X$  with respect to  $Y$ , and let  $\rho_{XX'}$  be the reliability of  $X$  at unit length. Show that if a squared validity  $\tilde{\rho}_{XY}^2$  is desired, then the length  $k$  of  $X$  must be

$$k = \frac{\tilde{\rho}_{XY}^2(1 - \rho_{XX'})}{\rho_{XY}^2 - \rho_{XX'} \tilde{\rho}_{XY}^2}.$$

- 5.20. Suppose that an examiner wishes to increase the validity of a test of unit length from 0.70 to 0.80 when the reliability of the test initially is 0.60. What must the length of the new test be?
- 5.21. Suppose the examiner wishes to increase the validity of this test to 0.95. Why does the formula in Exercise 5.19 give an absurd result?

### References and Selected Readings

- CRONBACH, L. J., and GOLDINE C. GLESER, The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, 1964, **24**, 467–480.
- FRENCH, J. W., The validity of new tests for the performance of college students with high-level aptitude. *Research Bulletin 63-7*. Princeton, N.J.: Educational Testing Service, 1963.
- GREEN, B. F. JR., A note on the calculation of weights for maximum battery reliability. *Psychometrika*, 1950, **15**, 57–61.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.
- MOSIER, C. I., On the reliability of weighted composites. *Psychometrika*, 1943, **8**, 161–168.
- NOVICK, M. R., The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, **3**, 1–18.
- PARZEN, E., *Stochastic processes*. San Francisco: Holden Day, 1962.
- PEEL, E. A., Prediction of a complex criterion and battery reliability. *British Journal of Psychology*, 1948, **1**, 84–94.
- THOMSON, G. H., Weighting for battery reliability and prediction. *British Journal of Psychology*, 1940, **30**, 357–366.
- WOODBURY, M., The stochastic model of mental testing theory and an application. *Psychometrika*, 1963, **28**, 391–394.
- WOODBURY, M., and F. M. LORD, The most reliable composite with a specified true score. *British Journal of Statistical Psychology*, 1956, **9**, 21–28.
- WOODBURY, M. A., and M. R. NOVICK, Maximizing the validity of a test battery as a function of relative test lengths for fixed total testing time. *Journal of Mathematical Psychology*, 1968, **5**, in press.

## CHAPTER 6

# FACTORS AFFECTING MEASUREMENT PRECISION, ESTIMATION, AND PREDICTION

### 6.1 Introduction

Chapters 2 through 5 were devoted to the definition of the basic constructs of mental test theory and the definition and exploration of the classical test theory model. In Chapters 7 through 9, we shall deal with methods for estimating values descriptive of these constructs, using both the classical model and a natural extension of it. However, before we derive these formal estimation procedures, it will be useful to consider in some detail the various factors that affect the interpretation of these constructs. An appraisal of these factors leads to specific practical restrictions on the experimental designs used for estimating the values of these constructs. Furthermore it will be convenient to introduce some simple experimental paradigms which may be used to define and estimate various coefficients of measurement precision. This will then lead to a consideration of the limitations on the applications of many of the formulas of Chapter 3.

### 6.2 Effect of Group Heterogeneity on Test Reliability

It is well known that the size of a correlation coefficient depends very much on the nature of the population in which measurements are made. A typical scatterplot of observed and true scores of a test shows a concentration of scores from the lower left to the upper right, indicating a substantial correlation between  $X$  and  $T$  and hence at least a moderate reliability for  $X$ . But if the scatterplot is limited to some small interval  $\tau_0 < T < \tau_1$ , there usually is no marked concentration of points from lower left to upper right, and hence there is at most a negligible correlation between  $X$  and  $T$ .

We can demonstrate this fact concretely either by surveying models based on specific distributional assumptions or by noting the following distribution-free result.

**Theorem 6.2.1.** Let  $A$  and  $A'$  be populations of examinees such that  $A'$  is contained in  $A$ . Suppose that for measurement  $X$  defined on  $A$ , the errors are homoscedastic, that is,

$$\sigma^2(E | \tau) = \sigma_E^2.$$

Let  $\tilde{X}$  be the measurement  $X$  restricted to  $A'$ . Assume that the restriction from  $A$  to  $A'$  is made by deleting at random persons with true scores in some specified set of  $\tau$  values. Then the reliability in the restricted population is given by

$$\rho_{\tilde{X}\tilde{\tau}}^2 = 1 - \frac{\sigma_X^2}{\sigma_{\tilde{X}}^2} (1 - \rho_{X\tau}^2).$$

*Proof.* From (3.3.8), we have

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{X\tau}^2).$$

But since the selection of persons is made on the basis of their  $\tau$ -values only, and since  $\sigma^2(E | \tau)$  is constant for all  $\tau$ , it follows that  $\sigma^2(\tilde{E}) = \sigma^2(E)$  for any such subpopulation, so that

$$\sigma_E^2 = \sigma_{\tilde{E}}^2 = \sigma_{\tilde{X}}^2(1 - \rho_{\tilde{X}\tilde{\tau}}^2) \quad \text{and} \quad \sigma_X^2(1 - \rho_{X\tau}^2) = \sigma_{\tilde{X}}^2(1 - \rho_{\tilde{X}\tilde{\tau}}^2).$$

Hence

$$\rho_{\tilde{X}\tilde{\tau}}^2 = 1 - \frac{\sigma_X^2}{\sigma_{\tilde{X}}^2} (1 - \rho_{X\tau}^2). \quad (6.2.1)$$

We see that  $\rho_{\tilde{X}\tilde{\tau}}^2$  is a strictly increasing function of  $\sigma_{\tilde{X}}^2/\sigma_X^2$  and that  $\rho_{\tilde{X}\tilde{\tau}}^2 = \rho_{X\tau}^2$  when  $\sigma_{\tilde{X}}^2 = \sigma_X^2$ . Thus  $\rho_{\tilde{X}\tilde{\tau}}^2$  is greater than or less than  $\rho_{XX'}$  according to whether  $\sigma_{\tilde{X}}^2$  is greater than or less than  $\sigma_X^2$ . When we thus restrict the population  $A$  to  $A'$  and thereby reduce the observed score variance, we also decrease the reliability coefficient associated with the measurement procedure. We might further note that in fact the derivation in no way depends on the assumption that  $A'$  is contained in  $A$ . Indeed we might have  $A$  contained in  $A'$  and  $\sigma_{\tilde{X}}^2 > \sigma_X^2$ , and Theorem 6.2.1 would still be valid.

Since  $0 < \rho_{X\tau}^2 < 1$ ,  $\sigma_X^2 = \sigma_{\tau}^2 + \sigma_E^2$ , and  $\sigma_{\tilde{X}}^2 = \sigma_{\tilde{\tau}}^2 + \sigma_E^2$ , the reader may easily verify that

$$0 < (\sigma_X^2/\sigma_{\tilde{X}}^2)(1 - \rho_{X\tau}^2) < 1.$$

Hence, when appropriate values are substituted for  $\sigma_X^2$ ,  $\sigma_{\tilde{X}}^2$ , and  $\rho_{X\tau}^2$ , formula (6.2.1) must yield a value  $0 < \rho_{\tilde{X}\tilde{\tau}}^2 < 1$ . For fixed  $\sigma_X^2$  and  $\rho_{XX'}$ , it is clear that  $\rho_{\tilde{X}\tilde{\tau}}^2$  is a strictly increasing function of  $\sigma_{\tilde{X}}^2$ .

Some further insight into this *correction for restriction of range* can be obtained from the following analysis: In many applications the data are such that

$$\sigma_{\tilde{E}}^2 \leq \sigma_E^2, \quad (6.2.2)$$

that is, the average error variance in the restricted population  $A'$  is less than or equal to the average error variance in the unrestricted population  $A$ . This may occur, for example, if the errors have approximately a binomial distribution

(see Chapter 23) and only the upper 10 or 15%, say, of the examinees are *included* in the restricted group. For such models we often have  $\sigma_E^2 \geq \sigma_{\tilde{E}}^2$ , or

$$\rho_{X_T}^2 \leq 1 - \frac{\sigma_{\tilde{X}}^2}{\sigma_X^2} (1 - \rho_{\tilde{X}\tilde{T}}^2), \quad (6.2.3)$$

where equality is attained if and only if  $\sigma_{\tilde{E}}^2 = \sigma_E^2$ .

The effect of this analysis is to show that a straightforward application of (6.2.1) to determine the reliability of a test in the unselected population yields a value which, under the condition (6.2.2), is too high. This value will be high to the extent that the error variance in the unselected population is higher than that in the selected population. On the other hand, in many applications the determination of the reliability in a restricted group from the reliability in an unrestricted group reverses the inequalities in (6.2.3) and provides a lower bound on the reliability in the restricted group. With a binomial error model, this would typically occur if only the lower 10 or 15%, say, of examinees were excluded from the restricted group. It should be noted that although it is necessary to assume  $\sigma^2(E | \tau) = \sigma_E^2$  for Theorem 6.2.1 to be true for every pair of populations, so strong an assumption is not required in any single application. If we are interested only in the one pair of populations at hand, we require only that the error variances be equal in these populations.

An example of the error that may occur from the misapplication of Theorem 6.2.1 will be of interest. Suppose test  $B$  has a reliability of 0.7 in the *unselected* population and test  $A$  has a reliability of 0.5 in the *selected* population, and that  $\sigma_X^2/\sigma_{\tilde{X}}^2 = 2$  for both tests. If we were to assume that  $\sigma_E^2 = \sigma_{\tilde{E}}^2$ , we would conclude that the reliability of test  $A$  in the unselected population is 0.75, a value *higher* than that for test  $B$ . However, if in fact  $\sigma_E^2 = 1.5 \sigma_{\tilde{E}}^2$  (which in some circumstances might be a more plausible assumption), then the reliability of test  $A$  in the unselected population would be 0.63, a value *lower* than that for test  $B$ .

Peters and Van Voorhis (1940) and Gulliksen (1950a) have surveyed some of the empirical results and critical comments that are pertinent to (6.2.1). These results and comments, we feel, justify the somewhat cautious use of Theorem 6.2.1 that we have recommended. A recent study by Boldt (1966) provides fresh evidence against the assumption of homogeneity of error variances. More such studies are needed.

### 6.3 Speed and Power Tests

We have previously indicated that the terms *test* and *measurement* do not have identical meanings, despite the fact that we ordinarily, in quite casual fashion, use these words interchangeably. The need for this distinction is particularly clear, and its importance in the selection and proper use of formulas is particularly apparent when we consider the distinction between *speed* and *power* tests.

If we take a person's *total incorrect score*  $I$  to be the total number of items that he does not answer correctly, it is clear that we may think of this score as the sum of a score  $W$ , the number of items to which he gives wrong answers, and  $U$ , the number of items that he does not attempt. Thus we write

$$I = W + U. \quad (6.3.1)$$

Since the test is timed, it may happen that a student cannot complete the test and that therefore his  $U$ -score is necessarily positive.

We now distinguish two "types" of measurements. If all students are given all the time that is necessary for them to complete all the items, and they in fact do so, then the random variable  $U$  may take only the value 0 [implying  $\mathcal{E}U = 0$  and  $\sigma^2(U) = 0$ ] and, hence,  $I = W$ . A test administered in this way is called a *power test*. Note that whether or not a test is a power test depends not only on the content of the test but on the conditions (in this case, the time limit) under which the test is administered.

At the other extreme, suppose that all items of a test are such that every subject who attempts an item answers it correctly. Because different students work at different rates of speed, however, they will tend to complete a different number of items in any fixed time period. Further, suppose that for the time period specified for the test, there are enough items so that no subject is able to complete all the items. In this case, the random variable  $W$  may take only the value zero [implying  $\mathcal{E}W = 0$  and  $\sigma^2(W) = 0$ ] and, hence,  $I = U$ . Such a test is called a *speed test*.

Pure speed and pure power tests are rarely employed in practice. Rather, most tests are what may be called *partially speeded* tests, the degree of speededness depending largely on the amount of time allotted for the test. These facts have relevance to the ways in which we may legitimately employ the various formulas developed in preceding chapters. There is no *theoretical* problem here so long as we recognize that these formulas refer to measurements and that the definition of a measurement describes much more than the item content of the test. In particular, it is important to recognize that in some cases the length of a test must be defined in terms of number of items, and in other cases the length of a test must be defined in terms of the actual time interval permitted for testing. Some reliability theory for speeded tests is given by Guttman (1955). The *practical* difficulties introduced by the speed factor will be discussed in some detail in succeeding sections.

In the present section we shall give a theoretical analysis of speed tests, due primarily to Gulliksen (1950a, 1950b). Although this analysis is of considerable conceptual interest, it does not lead to any particularly useful formulas.

First, we write

$$\sigma_x^2 = \sigma_W^2 + \sigma_U^2 + 2\rho_{WU}\sigma_W\sigma_U. \quad (6.3.2)$$

For a pure speed test,

$$\sigma_x^2 = \sigma_U^2, \quad (6.3.2a)$$

and for a pure power test,

$$\sigma_X^2 = \sigma_W^2. \quad (6.3.2b)$$

Gulliksen (1950a, 1950b) has suggested that if  $\sigma_W^2/\sigma_X^2$  is very small, then the test is primarily a speed test, and if  $\sigma_U^2/\sigma_X^2$  is very small, then the test is primarily a power test. The advantage of this index is that we can estimate its value from a single administration of the test.

Cronbach and Warrington (1951) have suggested another index of speededness. Let  $\rho_{SP}$  be the correlation between a test given as a speed test (i.e., with a fixed time limit) and the same test given as a power test (i.e., without a fixed time limit). Let  $\rho_{SS'}$  and  $\rho_{PP'}$  be the reliabilities of the test when given as a speed test and when given as a power test. Then

$$\rho_{SP}/\sqrt{\rho_{SS'}\rho_{PP'}}$$

is the disattenuated correlation between the speed and power measurements. The quantity

$$1 - \frac{\rho_{SP}^2}{\rho_{SS'}\rho_{PP'}}, \quad (6.3.3)$$

the percentage of variance attributable to speededness, thus seems to be a natural *index of speededness*. The following experimental paradigm suggested by Cronbach and Warrington may be used to estimate this index.

First, construct parallel forms *A* and *B*. Administer *A* with a time limit and, for each examinee, mark the last item attempted within the time limit. Then have the examinee review his answers and complete the test without time limit, using a different-colored pencil. Speed and power scores  $S_A$  and  $P_A$  can then be obtained for each examinee. Next, after some appropriate time interval, administer form *B* in the same manner. Speed and power scores  $S_B$  and  $P_B$  can then be obtained for each examinee. The sample correlations  $r(S_A, P_B)$  and  $r(S_B, P_A)$  are reasonable estimates of  $\rho_{SP}$ , and  $r(S_A, S_B)$  and  $r(P_A, P_B)$  are reasonable estimates of  $\rho_{SS'}$  and  $\rho_{PP'}$ . Then

$$1 - \frac{r(S_A, P_B)r(S_B, P_A)}{r(S_A, S_B)r(P_A, P_B)} \quad (6.3.4)$$

may be taken as a reasonable estimate of (6.3.3). Equation (6.3.4) is the form given by Cronbach and Warrington.

#### 6.4 Conditions of Measurement Affecting Reliability

In Chapter 2, we showed that the reliability of a measurement is equal to the correlation between parallel measurements. But explicit in the definition of parallel measurements is the requirement that the true score of each person be the same on each measurement. In practice, however, a person's responses to

parallel items or tests are affected, after a while, by practice, fatigue, and memory effects, and by actual (and not ephemeral) changes in his ability level.

In Chapters 7 and 9, we shall show that the standard procedures for estimating the reliability coefficient  $\rho(X_{g*}, X_{g'*})$  and the (average) error variance  $\sigma^2(E_{g*})$  require at least two measurements on each of a number of persons, with *the accuracy of the estimates increasing with the number of persons measured*. In contrast, the estimation of the error variance  $\sigma^2(E_{ga})$  for a person  $a$  requires at least two measurements on a person  $a$ , with *the accuracy of the estimates increasing with the number of observations on the person (a) measured*. But while it may be reasonable to assume, under certain conditions, that two successive measurements on a person may be considered parallel, it is seldom true in practice that any great number of repeated measures approximate parallel measurements. It is partially for this reason that classical test theory has concentrated on groups rather than individuals. Thus classical test theory is concerned with estimating parameters like  $\rho(X_{g*}, X_{g'*})$ ,  $\sigma^2(T_{g*})$ , and  $\sigma^2(E_{g*})$ , and not parameters like  $\sigma^2(E_{ga})$ .

The correlation between truly parallel measurements taken in such a way that the person's true score does not change between them is often called *the coefficient of precision*. As a reliability coefficient, we may write it as

$$\rho_{XT}^2 = \rho_{XX'} = \sigma_T^2 / \sigma_X^2 = 1 - (\sigma_E^2 / \sigma_X^2). \quad (6.4.1)$$

For this coefficient, the only source contributing to error variance is the unreliability or imprecision of the measurement procedure. This is the variance ratio that would apply if a measurement were taken twice and if no practice, fatigue, memory, or other factor affected repeated measurement. In most practical situations, other sources of error variation affect the reliability of measurement, and hence the coefficient of precision is not the appropriate measure of reliability. However, this coefficient can be thought of as representing the extent to which test unreliability is due solely to inadequacies of the test form and testing procedure, rather than the extent to which it is due to changes in people over time and lack of equivalence of nearly parallel forms.

In practice, there are three general methods for estimating reliability with parallel measurements. These are the *test-retest method*, the *internal analysis method*, and the *parallel forms method*. In the *test-retest method*, the same test form is administered to each person twice. The correlation between these measurements may then be taken, with caution, as an approximation to the coefficient of precision. There are several problems with the test-retest method. First, there is likely to be a practice or memory effect on the second administration, so that errors on the two administrations tend to be correlated. Then there is the possibility that repeated testing may fatigue the examinee and thereby introduce a biased error which, of course, is absorbed into the true score. Also any temporary or permanent changes in the examinee's ability contribute to error.

When working with the *test-retest method*, it is often difficult to know whether, on balance, these effects are likely to inflate or deflate the correlation between the repeated measurements. According to Gulliksen (1950a), they usually tend to inflate the correlation, particularly if the period between repeated administrations is small and fatigue effects are minimized. Under these conditions, the test-retest method yields an overestimate of the coefficient of precision. Guttman (1955), however, is much more favorably disposed toward the retest method. Certainly, as the time interval between administrations increases, memory effects become less important and changes in the persons become more important, and these alterations result in a decrease in the test-retest correlation. One application in which the test-retest method seems particularly useful is in correcting for attenuation (see Section 6.5); here an overestimate of reliability results in a conservative correction for attenuation, which is usually desirable.

The second method of estimating reliability involves an internal analysis of the variances and covariances of the test items. For example, the test may be divided into two parts called *split halves*. We may do this by assigning the items randomly to the halves or by matching them in pairs on means and standard deviations and then randomly assigning one item from each pair to each half. Then, if we ignore for the present the problems of statistical estimation, the correlation between scores on the half-tests gives the reliability of a test of half length. Since we have matched the two halves on means and standard deviations, we may consider them to be approximately parallel and use the Spearman-Brown formula (5.10.1) to obtain the reliability of the whole test, the so-called *stepped-up reliability*. However, as indicated in Exercise 4.16, the Spearman-Brown formula will give too small a value if the splits are not, in fact, parallel.

If the halves can only be considered to be essentially  $\tau$ -equivalent, then, considering Eq. (4.4.4) as an equality, we can use coefficient  $\alpha$  to obtain the stepped-up reliability. If the parts are not essentially  $\tau$ -equivalent, coefficient  $\alpha$  will give a lower bound to the stepped-up reliability. We may employ split thirds, fourths, and so on in a similar manner, using (4.4.9) when the parts are parallel and (4.4.4) as an equality when the parts are essentially  $\tau$ -equivalent. If parallel splits can in fact be obtained, an appropriate application of the Spearman-Brown formula provides an excellent approximation to the reliability of the total test. However, it is often difficult (and expensive) to obtain such parallel splits in practice.

This scheme of dividing a test into component parts and then applying a formula for stepped-up reliability must be applied with caution. If the items of a test are placed in order of increasing difficulty, as is common, then the first half of the test is not at all comparable to the second half. But even when the items are not so ordered, the time limit of the test generally has a greater effect on the second-half scores than on the first-half scores because many students who finish the first half of the test do not finish the second half. In practice, the effect of this is to make the first half essentially a power test and the second

half at least partially a speeded test. When dealing with a partially speeded test, therefore, it is bad procedure to divide it into split halves by placing the first  $n/2$  items in the first component and the second  $n/2$  items in the second.

However, proper split halves can be obtained for homogeneous tests by dividing the total testing time in fractions and identifying the length of the test, not with the number of items, but with the time allotted for taking it. For example, consider a speeded test of one hour's duration that consists of items that would be homogeneous if the test were not speeded. We may divide this test into four separately timed tests, each of fifteen minutes' duration. Then, if we take these four subtests to be components, we can see that the composite must be homogeneous since each of its components is speeded to exactly the same degree. Therefore, although a test is a speeded one, split halves, thirds, and so on obtained in this way will be parallel, provided that the items are homogeneous and that no fatigue or practice effects are present.

Division of a test into component parts by placing alternate odd- and even-numbered items into the respective halves (the odd-even method) is a common and usually satisfactory approach. However, some care must be exercised even with this method. The most serious problem with the odd-even method arises with speeded or partially speeded tests. Any student failing to complete the last  $2r$  items on the test will have a score of zero on each of these items, and therefore the odd scores and the even scores for these items will be perfectly correlated. When these items are pooled with the items that the subject answered, the computed split-half reliability may be substantially different from the actual reliability of the test.

The reader can apply the basic principles developed here in his consideration of any proposed method of estimating reliability. He need only ask himself whether the necessary assumption of parallelism or  $\tau$ -equivalence of *measurements*, not tests, is reasonably satisfied. If they are not met, he should seek a more appropriate experimental design, if one is possible.

If the test is either reasonably homogeneous or of substantial length and not speeded, coefficient  $\alpha$  should provide a very usable approximation to the coefficient of precision. Cronbach and Azuma (1962) have thrown some light on the question of how homogeneous the test must be and/or how long it must be. When working with tests of moderate length that have been specifically designed to measure a single underlying trait, coefficient  $\alpha$  should provide a useful approximation to the reliability coefficient for most procedures that do not involve corrections for attenuation.

The third method of obtaining (approximately) parallel measurements is the *parallel forms method*. To employ this method we must construct two forms of a test, each consisting of items that are different but that provide (approximately) parallel measurements when the two forms are administered. Where we may suppose that additional intra-examinee variation is introduced because of actual changes in the examinee or variations in the examining conditions, we call the correlation between the successive measurements obtained from parallel

forms *a coefficient of stability* (Cronbach, 1947). Note that we speak of *a* coefficient of stability rather than *the* coefficient of stability; we do this because a coefficient of stability depends on the conditions of retesting, length of time interval, etc., and thus each test has many coefficients of stability.

A coefficient of stability may be related to equation (6.4.1). The error variance  $\sigma_E^2$  for this coefficient is composed of two components, an imprecision of the measuring procedure and also a variation between true scores on the two measurements on each person. Thus a coefficient of stability is always less than the coefficient of precision. [Cureton (1958) has given a different definition, in which the error variance does not include the effect of imprecision of measurement.] When the conditions of the two administrations are equivalent and the intervening time is short, the parallel forms method usually produces a correlation coefficient which is close to the coefficient of precision.

Quantities to which we usually refer as *coefficients of equivalence* and *coefficients of equivalence and stability* are also associated with the parallel forms method. The term "coefficient of equivalence" is used to indicate the correlation between nearly parallel forms. In terms of (6.4.1), the coefficient of equivalence treats as error both the imprecision of the measurement and also the true-score variance that is due to the lack of parallelism of the approximately parallel forms. The coefficient of equivalence can often be approximated by the successive administration of parallel forms under conditions that minimize fatigue and practice effects and conditions that minimize any change of true-score level of the persons being measured. Ordinarily the coefficient of equivalence is somewhat lower than the coefficient of precision. In fact, however, the administration of parallel forms leads to a coefficient of stability and equivalence which includes errors due to imprecision, variation between forms, and variation of an examinee's ability between administrations.

Much has been written on determining the "best" method of estimating reliability. Often such writing is in the form of authoritative pronouncements. Rather than add further (possibly conflicting) pronouncements, we shall simply encourage the reader to consider, in each situation he meets, the various factors that could affect the quantities he wishes to compute and then to judge, in accordance with the theories developed here, the extent to which these factors must be weighted. If each investigator proceeds on this basis and makes public his reasons for adopting a particular technique, we shall at least have an objective basis for evaluating each computation. We shall present further considerations of these problems within the context of a more general theory in Chapters 7 through 9 and 11.

## 6.5 Experimental Problems in Correcting for Attenuation

In Chapter 3 and again in Chapter 5, we considered so-called attenuation formulas, which give the correlation between measurements "corrected" for the unreliability of one or both measurements. One such formula, given as (3.9.6)

and also as (5.11.3), is

$$\rho(T_X, T_Y) = \rho_{XY} / \sqrt{\rho_{XX'}\rho_{YY'}} \quad (6.5.1)$$

where  $\rho_{XX'}$  is the reliability of  $X$ , etc. In Chapter 3, this result was obtained by evaluating each of the four correlations in terms of variances and covariances. In Chapter 5, this result was obtained by showing that the right-hand member of (6.5.1) is the limit, as  $k$  and  $l$  tend to infinity, of the relative observed scores  $X(k)/k$  and  $Y(l)/l$ :

$$\rho(T_X, T_Y) = \lim_{\substack{k \rightarrow \infty \\ l \rightarrow \infty}} \rho \left[ \frac{x(k)}{k}, \frac{y(l)}{l} \right] = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}}.$$

It would seem that a simple way to obtain the disattenuated correlation  $\rho(T_X, T_Y)$  is to estimate the quantities  $\rho_{XY}$ ,  $\rho_{XX'}$ , and  $\rho_{YY'}$  (from a sufficiently large sample, so that possible sampling fluctuations can be ignored) and then to use (6.5.1) to solve for  $\rho(T_X, T_Y)$ . Unfortunately this procedure is often unsatisfactory. Indeed, in some cases this approach has led to the impossible conclusion that  $\rho(T_X, T_Y) > 1$ . The discussion in the previous sections of this chapter, however, enables us to understand this paradox and to avoid incorrect inferences.

The reader should note that the quantities  $\rho_{XX'}$  and  $\rho_{YY'}$  appear in the denominator of (6.5.1). Hence, if these values are understated, the value of the ratio will be overstated. In practice, this is what all too often occurs when the correlation between  $X$  and  $Y$  is estimated on the basis of a single administration of each of the measurements  $X$  and  $Y$  in close temporal contiguity. In this case, the error variation present in the estimate is the error due to the imprecision of each of the two measurements.

On the other hand, estimates of  $\rho_{XX'}$  and  $\rho_{YY'}$  typically require at least two measurements on each  $X$  and  $Y$ . These repeated measurements are often obtained through test-retest or parallel-forms methods. However, *in each case at least one additional source of variation may be introduced which is not present in the estimation of  $\rho_{XY}$* . For example, if we use the parallel-forms method, then any differences in the (ostensibly) parallel forms will affect  $\rho_{XX'}$  and  $\rho_{YY'}$ , but not  $\rho_{XY}$ . In the language of previous sections, we might say that we require a coefficient of precision but usually (and incorrectly) use a coefficient of equivalence or a coefficient of stability. If we use an internal-analysis method, then a lack of homogeneity of the items (or subtests) will again result in an underestimation of reliability. Hence the correction for attenuation tends to be an overcorrection of unknown extent. If we are to use the correction for attenuation precisely, we must use it only in conjunction with an experimental design that assures that essentially no more or less error variation is introduced into the estimate of  $\rho_{XX'}$  and  $\rho_{YY'}$  than is introduced into the estimate of  $\rho_{XY}$ . The reader may profitably consult Garside (1958).

## 6.6 Accuracy of the Spearman-Brown Prophecy Formulas

Chapter 5 contains numerous formulas that express the error variances, reliabilities, and validities of a test of length  $k$  in terms of a test of unit length. The length of a test is defined there as the number of *parallel* components comprising the composite. If  $k$  parallel components could indeed be given, then, except for sampling fluctuations, these formulas would accurately forecast the characteristics of the lengthened test. But completely parallel components are not usually attainable unless one takes only a very small number of measurements. Rather, as  $k$  becomes large, the repeated measurements become increasingly less parallel, and nonequivalence and instability factors increase. As indicated in Exercise 4.16, if the parallelism is imperfect, the Spearman-Brown formula gives too small a value for the reliability of the whole test.

On the other hand, the possible violation of the assumption of experimental independence could result in an underestimate of the error variance of the lengthened test and therefore in an overestimate of its reliability and validity. There is no question that if this formula is used unwisely, the resulting value will have little relationship to reality. However, if the application of the formula is restricted to situations in which the assumptions are reasonably well satisfied, quite satisfactory results may be obtained.

## 6.7 Reliability as a Generic Concept

The discussion in the previous sections suggests that reliability is a generic concept referring to the accuracy (equivalence and stability) of measurement. We have shown how several different coefficients can be brought into the mathematical framework that defines reliability as the squared correlation  $\rho_{XT}^2$ . We did this by defining the measurement  $X$  to correspond with the experimental (measurement) procedure being used and treating as error various specific factors that affect this measurement. In Chapters 7 through 9, we shall see that a more formal and more satisfactory treatment of this subject can be given by defining *different* true scores, each appropriate to the particular reliability of interest. Also, from our discussion in Section 6.2 we know that it is equally important to specify the population in which the reliability is defined.

In the 1966 *Standards for Educational and Psychological Tests and Manuals*, the classification of reliability coefficients into several types (e.g., coefficient of precision and coefficient of stability) has been discarded almost entirely. It is pointed out there (p. 26) that "such a terminological system breaks down as more adequate statistical analyses are applied and methods are more adequately described". Instead it is recommended that "test authors work out suitable phrases to convey the meaning of whatever coefficients they report; as an example, the expression, 'the consistency between measurements by different test forms as determined by stability over a 7-day interval', although lengthy, will be reasonably free from ambiguity". This more flexible and more informative approach is based on the realization that many components contribute to

variations among observations and the understanding that the decision as to which of these components are and which are not to be treated as contributing to "error" depends on the particular use to which the coefficient is to be put. As the *Standards* point out,\*

There are various components that may contribute to inconsistency among observations: (a) response-variation by the subject (due to changes in physiological efficiency, or in such psychological factors as motivation, effort, or mood): these may be especially important in inventories of personality; (b) variations in test-content or the test-situation (in "situational tests" which include interacting persons as part of the situation, this source of variation can be relatively large); (c) variations in administration (either through variations in physical factors, such as temperature, noise, or apparatus-functioning, or in psychological factors, such as variation in the technique or skill of different test-administrators or raters); (d) variations in the process of observation. In addition to these errors of observation, scoring error variance in test scores reflects variation in the process of scoring responses as well as mistakes in recording, transferring, or reading of scores.

There have been many suggestions that the subject of reliability can be studied more accurately by considering stochastic process models that explicitly take into account such factors as practice, fatigue, and specific learning. A good example of this kind of model is that suggested by Hoffman (1963). This model assumes linear homogeneous gains. Undoubtedly such models, if accurate, would help us understand and predict the behavior of examinees on repeated testings. At the present time, however, no widely applicable, detailed stochastic model yielding experimentally verified results has been stated. Until this is done we can only speculate on the potential contribution of this kind of model to the study of reliability. A step in this direction, however, has been taken by Vinsonhaler and Meredith (1966), who correctly place their emphasis on the reliability of the first observation. A brief survey of an important part of the literature and some experimental results on retesting may be found in Howard (1964).

### 6.8 Effect of Explicit and Incidental Selection on Test Validity: The Two-Variable Case†

In Section 6.2, we demonstrated that the interpretation of reliability coefficients depends strongly on the heterogeneity of the group. In this and the next three sections, we shall study the effect of group heterogeneity on validity, following at first the distinction between explicit and incidental selection em-

\* From French, J. W., and W. B. Michael, Standards for Educational and Psychological Tests and Manuals, Copyright, 1966. Washington, D.C.: *American Psychological Association*. Used by permission.

† Reading of this and the following two sections may be postponed, if desired, and undertaken in conjunction with the reading of Chapter 12.

phasized by Gulliksen (1950a). Later we shall see, as Gulliksen himself has indicated, that this distinction is not fundamental.

Whenever two tests are compared and only one of them has been used in selection, then we must take this fact into account if the comparison is to be valid. We particularly emphasize this point because it is seldom, if ever, the case that we encounter a group on which criterion data are available, the members of which have not been subject to selection.

If some measurement is used as the basis for selection, then we say that there has been *explicit selection* on that variable, and we call this variable that has been subject to explicit selection the *explicit selection variable*. A typical example is the use of an entrance examination to select students from an *applicant group* to form a *selected group* for admission to college. In some situations the selected group is obtained simply by selecting each applicant whose observed test score is equal to or greater than some specified number. (More typically, and more appropriately, other factors are used in the selection process.) Such selection obviously reduces the variance of the predictor variable in the selected group as compared with the applicant group. For example, suppose that test scores in some group are normally distributed with unit variance, and suppose that the top  $p\%$  of the group is taken as the selected group. Then the standard deviation in the selected group will be as given in Table 6.8.1. Although the normal distribution is not a particularly good test score model for many purposes, the standard deviations listed in Table 6.8.1 should give reasonable approximations to the values that may be encountered in practice, provided that the data are symmetrically distributed.

In addition to this reduction in the standard deviation of the selection variable, there will be a reduction in the variance in the selected group of any variable

**Table 6.8.1**  
Variance in the selected group when the top  $p$  percent  
are selected from a standard normal population

Normal deviate $Z$	$-\infty$	-1.25	-0.85	-0.50	-0.25	0	0.25	0.50	0.85	1.25
Percent selected	100	89.44	80.23	69.15	59.87	50.00	40.13	30.85	19.77	10.56
Standard deviation in the selected group	1	0.84	0.76	0.70	0.65	0.60	0.56	0.52	0.47	0.41
Ratio of standard deviation in unselected group to selected group	1	1.19	1.31	1.43	1.54	1.64	1.79	1.93	2.14	2.42

correlated with the selection variable. Thus, to the extent that a correlation exists between this second variable and the explicit selection variable, we can say that there has been an indirect selection on this second variable. We refer to this second kind of selection as *incidental selection* and refer to these variables as *incidental selection variables*. Such indirect selection also results in a lowering of the validity coefficient for the second variable, though the amount of decrease is typically less than for explicit selection.

Explicit and incidental selection affect not only the absolute size of the validity coefficients but also their relative size (Thorndike, 1949), "so that the test which is really most valid as applied to the general run of applicants may appear to be one of the less valid in a group resulting from high standards of preselection". Thorndike reports the following results, obtained from a study conducted as part of the AAF Aviation Psychology Program:

Predictor variable	Validity coefficients	
	Total group <i>N</i> = 1036	Qualified group <i>N</i> = 136
Pilot Stanine (Composite Score)	.64	.18
Mechanical Principles Test	.44	.03
General Information Test	.46	.20
Complex Coordination Test	.40	-.03
Instrument Comprehension Test	.45	.27
Arithmetic Reasoning Test	.27	.18
Finger Dexterity Test	.18	.00

Selection was on *Pilot Stanine Score*. The next four tests entered heavily into the Stanine Score. The last two tests had no weight in the Stanine Score. Clearly the "relative size of the validity coefficients in the restricted group gives little basis for judging the validity of the tests as applied to the complete population of applicants".

We first consider the bivariate situation where there has been explicit selection on  $X$  and where we are interested in the correlation between  $X$  and  $Z$  in the unselected group. The available data in the selected group are the correlation and both variances, and in the unselected group only the variance of the variable subject to selection. The simple selection situation where  $X$  is the selection (predictor) variable and  $Z$  is the criterion variable is one such case.

Let  $X$  and  $Z$  be the variables before selection and  $X$  the variable subject to explicit selection. Let  $X^*$  and  $Z^*$  be the variables after explicit selection on  $X$ . To obtain the "standard results" we require the following two assumptions:

**Assumption 6.8.1.** The true regression function of  $Z$  on  $X$  is a linear function of  $x$  throughout the domain of  $X$ :

$$\mathcal{E}(Z | x) = \alpha + \beta x. \quad (6.8.1)$$

**Assumption 6.8.2.** The conditional variance of  $Z$ , given  $x$ , does not depend on  $X$ :

$$\sigma^2(Z | x) = \sigma^2(Z | x'), \quad \text{for all } x, x'. \quad (6.8.2)$$

Since the true regression of  $Z$  on  $x$  is linear, the minimum mean squared error regression function is not affected by explicit selection on  $x$ ; hence the regression coefficients will be the same in the unselected group and in any selected group. Symbolically we write

$$\beta(Z | x) \equiv \rho(X, Z) \frac{\sigma(Z)}{\sigma(X)} = \rho(X^*, Z^*) \frac{\sigma(Z^*)}{\sigma(X^*)} \equiv \beta(Z^* | x^*). \quad (6.8.3)$$

Since the true regression is linear, Assumption 6.8.2 implies that the error of prediction from the minimum mean squared error regression line does not depend on  $x$  and hence will be the same in the selected and in the unselected groups. The equality of the residual variances in the unselected and selected populations may be expressed as

$$\sigma^2(Z | x) = \sigma^2(Z^* | x^*). \quad (6.8.4)$$

Under the assumed selection procedure, Equations (6.8.3) and (6.8.4) are jointly equivalent to (6.8.1) and (6.8.2) and represent the "standard" assumptions for this problem. The required results for the case introduced above are contained in

**Theorem 6.8.3.** Under Assumptions 6.8.1 and 6.8.2,

$$\rho_{XZ}^2 = \left[ 1 + \frac{\sigma_{X^*}^2}{\sigma_X^2} \left( \frac{1}{\rho_{X^*Z^*}^2} - 1 \right) \right]^{-1} \quad (6.8.5)$$

and

$$\sigma_Z^2 = \sigma_{Z^*}^2 [1 - \rho_{X^*Z^*}^2 + \rho_{X^*Z^*}^2 (\sigma_X^2 / \sigma_{X^*}^2)]. \quad (6.8.6)$$

*Proof.* Solving (6.8.3) for  $\sigma_Z$ , we have

$$\sigma_Z = \frac{\rho_{X^*Z^*}\sigma_{Z^*}\sigma_X}{\rho_{XZ}\sigma_{X^*}}. \quad (6.8.7)$$

From (6.8.4) and the usual formula for the residual variance (see 12.3.4),

$$\sigma_{Z^*}^2 (1 - \rho_{X^*Z^*}^2) = \sigma_Z^2 (1 - \rho_{XZ}^2).$$

Substituting for  $\sigma_Z^2$ , we have

$$\sigma_{Z^*}^2 (1 - \rho_{X^*Z^*}^2) = \frac{\rho_{X^*Z^*}^2 \sigma_{Z^*}^2 \sigma_X^2}{\rho_{XZ}^2 \sigma_{X^*}^2} (1 - \rho_{XZ}^2).$$

Then the first part of the theorem follows on solving for  $\rho_{XZ}^2$ . If this value is then substituted in (6.8.7), the second part of the theorem follows.  $\square$

Suppose the assumptions of the theorem hold and we observe a correlation  $\rho_{X^*Z^*} = 0.40$  in a selected population in which the variance of  $X^*$  is 80. Suppose we know that in the unselected population the variance of  $X$  is 100. Then, from (6.8.5), we compute the square of the correlation between  $X$  and  $Z$  in the unselected population to be

$$\rho_{XZ}^2 = \frac{1}{1 + \frac{80}{100(0.16)} - 1} = 0.1923$$

and the correlation to be 0.44. Now suppose that a more extreme selection is performed on the unselected group so that the variance in the selected group is 32. If the population distribution in the unselected group was approximately normal, a selection of about the upper 40% would be required. In this case, the correlation in the new restricted group would be only 0.24.

It should be noted that nowhere in the derivation is it necessary to employ the assumption that the selection is made explicitly on  $X$ . All that is required to prove the result of Theorem 6.8.3 is that the regression of  $Z$  on  $x$  be the same in the selected and unselected groups and that the variance of  $Z$  be the same in both groups. If these assumptions are satisfied, then the theorem is valid for values  $\sigma(X) < \sigma(X^*)$  as well as for values  $\sigma(X) \geq \sigma(X^*)$ , and regardless of how the selection is in fact made.

However, if selection has not been made explicitly on  $X$ , this assumption of linearity becomes very difficult to justify. It is ordinarily a serious error to ignore the fact that the selected group has been selected on some (usually unknown) variable other than  $X$  and to act as if the regression of  $Z$  on  $x$  were linear. In many applications this will not even be approximately true and serious errors will occur. Linn (1967) has convincingly demonstrated the seriousness of this problem.

### 6.9 The Effect of Selection on Test Validity: The Three-Variable Case

Suppose  $X$  is a predictor variable being used for selection,  $Z$  is the criterion variable, and  $Y$  is a proposed predictor. We desire to compare the relative merits of  $X$  and  $Y$  as predictors, but we are unable to discontinue selection on  $X$  in order to gather data for the comparison. The explicit selection on  $X$  has resulted in an incidental selection on both  $Y$  and  $Z$ . The important computation, which cannot be made from the theory of the previous section, is the correlation between  $Y$  and  $Z$  in the unselected group. If the required assumptions are satisfied, then the result of the following theorem may be used to make the necessary computation.

**Theorem 6.9.1.** Let  $X$  be a variable subject to explicit selection and let  $Y$  and  $Z$  be correlated with  $X$ , so that  $Y$  and  $Z$  are thus subject to incidental selection. Assume that the true regressions of  $Y$  on  $x$  and  $Z$  on  $x$  are linear. Further assume that the variance of  $Y$  given  $x$ , the variance of  $Z$  given  $x$ ,

and the covariance of  $Z$  and  $Y$  given  $x$ , do not depend on  $x$ . Then

$$\rho_{YZ} = \frac{[\rho_{Y^*Z^*} - \rho_{X^*Y^*}\rho_{X^*Z^*} + \rho_{X^*Y^*}\rho_{X^*Z^*}(\sigma_X^2/\sigma_{X^*}^2)]}{[1 - \rho_{X^*Y^*}^2 + \rho_{X^*Y^*}^2(\sigma_X^2/\sigma_{X^*}^2)]^{1/2}[1 - \rho_{X^*Z^*}^2 + \rho_{X^*Z^*}^2(\sigma_X^2/\sigma_{X^*}^2)]^{1/2}}, \quad (6.9.1)$$

where the star superscript again refers to the selected group.

*Proof.* The stated assumptions yield the equations

$$\rho_{X^*Y^*}\frac{\sigma_{Y^*}}{\sigma_{X^*}} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}, \quad (6.9.2)$$

$$\rho_{X^*Z^*}\frac{\sigma_{Z^*}}{\sigma_{X^*}} = \rho_{XZ}\frac{\sigma_Z}{\sigma_X}, \quad (6.9.3)$$

$$\sigma_{Y^*}^2(1 - \rho_{X^*Y^*}^2) = \sigma_Y^2(1 - \rho_{XY}^2), \quad (6.9.4)$$

$$\sigma_{Z^*}^2(1 - \rho_{X^*Z^*}^2) = \sigma_Z^2(1 - \rho_{XZ}^2), \quad (6.9.5)$$

$$\frac{\rho_{Y^*Z^*} - \rho_{X^*Y^*}\rho_{X^*Z^*}}{[(1 - \rho_{X^*Y^*}^2)(1 - \rho_{X^*Z^*}^2)]^{1/2}} = \frac{\rho_{YZ} - \rho_{XY}\rho_{XZ}}{[(1 - \rho_{XY}^2)(1 - \rho_{XZ}^2)]^{1/2}}. \quad (6.9.6)$$

Equation (6.9.6) states that the partial correlation of  $Y$  and  $Z$  does not depend on  $x$ . This follows from the assumption that the covariance of  $Y$  and  $Z$  and the partial variances of  $Y$  and  $Z$  do not depend on  $x$ . The proof then follows upon simple algebraic manipulation: Equations (6.9.2) through (6.9.5) are just those of Theorem 6.8.3; hence, from Eq. (6.8.5), we have

$$\rho_{XY}^2 = \left[ 1 + \frac{\sigma_{X^*}^2}{\sigma_X^2} \left( \frac{1}{\rho_{X^*Y^*}^2} - 1 \right) \right]^{-1}, \quad (6.9.7)$$

$$\rho_{XZ}^2 = \left[ 1 + \frac{\sigma_{X^*}^2}{\sigma_X^2} \left( \frac{1}{\rho_{X^*Z^*}^2} - 1 \right) \right]^{-1}. \quad (6.9.8)$$

Solving (6.9.6) for  $\rho_{YZ}$  and substituting for  $\rho_{XY}^2$  and  $\rho_{XZ}^2$  produces the stated result.  $\square$

Solving (6.9.4) for  $\sigma_Y^2$  and substituting for  $\rho_{XY}^2$  from (6.9.7) also gives

$$\sigma_Y^2 = \sigma_{Y^*}^2 \left[ \frac{\sigma_X^2}{\sigma_{X^*}^2} \rho_{X^*Y^*}^2 + (1 - \rho_{X^*Y^*}^2) \right]. \quad (6.9.9)$$

When  $\sigma_Y^2$  and  $\sigma_{Y^*}^2$  are available this equation may be useful as a partial test of the model.

Suppose that the correlation between a quantitative aptitude test  $X$  and statistics course grade  $Z$  is 0.22 in a group selected on  $X$ , and that the correlation of a verbal ability test  $Y$  and this course grade is 0.23. Suppose that the intercorrelation of these two tests in the selected group is 0.20. Further, suppose

that the variances of the mathematics test before and after selection are 100 and 25, respectively. This corresponds to a ratio of standard deviations of 2. If the explicit selection variable is approximately normal, then a selection of about the top 30% is indicated. Then in the unselected group the correlations are approximately 0.41 and 0.31, respectively, and the intercorrelation is approximately 0.38, as determined by the following computations:

$$\rho_{XZ} = [1 + \frac{1}{4}(0.0484 - 1)]^{-1/2} \doteq 0.41,$$

$$\rho_{YZ} = \frac{[(0.23) - (0.20)(0.22) + (0.20)(0.22)4]}{[1 - (0.20)^2 + (0.20)^2(4)]^{1/2}[1 - (0.22)^2 + (0.22)^2(4)]^{1/2}} \doteq 0.31,$$

$$\rho_{XY} = [1 + \frac{1}{4}(0.0400 - 1)]^{-1/2} \doteq 0.38.$$

Thus, as we would expect, the quantitative aptitude test is a substantially better predictor of performance in statistics courses than is the verbal aptitude test. The relatively low correlation in the unselected group might in part reflect the fact that a self-selection process has already restricted the range of the applicant group.

If we suppose that the second test is indeed a second quantitative aptitude test being offered as a replacement for the quantitative aptitude test currently in use, a second important application of the theory becomes evident. From the above example we see that if one is ignorant of the theory, he would be led to discard the old test in favor of the new one, while, if he knows the theory and applies it, he will surely retain the old test.

### 6.10 The Effect of Selection on Test Validity: The General Case\*

Lawley (1943) has provided a compact generalization of the result of the previous section to the case where a number of variables are subject to explicit selection and a further number of variables are correlated with these and therefore subject to incidental selection. The assumptions of Lawley's theorem are that the regression of each incidental selection variable on any set of explicit selection variables is linear and that given the explicit selection variables, the conditional variance-covariance matrix of the incidental selection variables does not depend on the particular values of the explicit selection variables. Lawley shows that these assumptions are necessary and sufficient for the establishment of the selection formulas.

Let  $\mathbf{X}$  be the  $p$ -element vector of explicit selection variables and  $\mathbf{Y}$  the  $(n - p)$ -element vector of incidental selection variables in the unselected group. Let  $\mathbf{V}$  be the variance-covariance matrix of  $\mathbf{X}^*$ ,  $\mathbf{Y}^*$ , where  $\mathbf{X}^*$  and  $\mathbf{Y}^*$  are random variables corresponding to  $\mathbf{X}$  and  $\mathbf{Y}$ , but in the selected rather than the applicant group. Partitioning  $\mathbf{V}$ , we write

$$\mathbf{V} = \begin{vmatrix} \mathbf{V}_{pp} & \mathbf{V}_{p,n-p} \\ \mathbf{V}_{n-p,p} & \mathbf{V}_{n-p,n-p} \end{vmatrix}.$$

---

\* Reading of this section may be omitted without loss of continuity.

Let  $W_{pp}$  be the variance-covariance matrix of  $\mathbf{X}$  in the unselected group. Then

**Theorem 6.10.1.** Under the above assumptions, the variance-covariance matrix of  $\mathbf{X}'$ ,  $\mathbf{Y}'$  in the unselected group is

$$\begin{vmatrix} \mathbf{W}_{pp} & \mathbf{W}_{pp}\mathbf{V}_{pp}^{-1}\mathbf{V}_{p,n-p} \\ \mathbf{V}_{n-p}\mathbf{V}_{pp}^{-1}\mathbf{W}_{pp} & \mathbf{V}_{n-p,n-p} - \mathbf{V}_{n-p,p}(\mathbf{V}_{pp}^{-1} - \mathbf{V}_{pp}^{-1}\mathbf{W}_{pp}\mathbf{V}_{pp}^{-1})\mathbf{V}_{p,n-p} \end{vmatrix}.$$

Lawley's elegant proof of this theorem is based on the use of moment generating functions, though no distributional assumptions are made. Lawley further points out that "... the selection formulae, if once applicable, may again be applied when a second selection is performed on the already selected population ... the operations  $\mathbf{W}_{pp} \rightarrow \mathbf{V}_{pp}$  and  $\mathbf{V}_{pp} \rightarrow \mathbf{U}_{pp}$  when performed in succession are equivalent to the single operation  $\mathbf{W}_{pp} \rightarrow \mathbf{U}_{pp}$ ".

Actually the method of selection plays no part in Lawley's formulation. The only requirements for the theorem are the linearity of regression of the second set of variables on the first and the homoscedasticity of the variance-covariance matrix of the second set of variables, given the first set. Explicit selection on the first set of variables by any means will preserve the required linearity and homoscedasticity.

A complete discussion of the many cases arising in practice that are covered by the general formulation of this section has been given by Rydberg (1963, now out of print). The most common applications are those in which complete information is available on the explicit selection variables and partial information is available on the implicit selection variables, as in Section 6.9. Solutions are also possible when certain combinations of partial information on the explicit and incidental selection variables are available. For many of these cases the number of explicit selection variables must be less than or equal to the number of implicit selection variables. Such cases are less common than those described in this chapter.

## 6.11 Accuracy of the Selection Formulas

Despite the obvious importance of this problem, relatively little work has been done to check on the accuracy of these formulas. A small study was undertaken in the AAF Aviation Psychology Program and reported by Thorndike (1949). Another has been reported by Peters and Van Voorhis (1940), and some by Rydberg (1963). Each of these writers takes a rather positive attitude toward the selection formulas. They suggest, on the basis of the studies they have seen, that the corrected values are more accurate than the uncorrected values and that the common tendency is for the formulas to undercorrect.

However, none of these studies involves extreme selection. The present writers feel that a more cautious attitude toward these formulas is called for in any applications in which the ratio of standard deviations in the unselected group to standard deviations in the selected group is more than 1.40. This condition corresponds to a selection of approximately the upper 70% from a

standard normal population. Unfortunately, in many applications the percentage selected is much lower than this and hence these applications of the theory should be questioned.

Three factors contribute to affect the accuracy of the selection formulas in many applications. Two of these factors are the failure of most test score data to satisfy the two assumptions of linearity and homoscedasticity of the general Theorem 6.10.1. We shall discuss one of these points in detail in Chapter 23, where we shall suggest that some binomial models typically give better fits to test score data than do normal distribution models, which imply linearity and homoscedasticity. A feature of the binomial models is that they do not satisfy the homoscedasticity assumption (6.8.2); rather, the variance in the tails of these distributions is often smaller than in the middle of the distributions. As with the correction of the reliability coefficient for restriction of range, this inaccuracy of the model tends to inflate the correction.

It is also true that for many bivariate test score data the regression function is not very accurately approximated in the tails of the distribution by the same line that approximates the regression function in the center of the distribution. Very often the regression slope is reduced in the upper tail. Thus, when selection is by lower truncation, Assumption 6.8.1 is not likely to be satisfied, even approximately. This inaccuracy of the model tends to deflate the correction. To some extent these two factors tend to cancel each other out.

As we have noted, a serious obstacle arises in most applications because selection is seldom made in the simple manner suggested in the previous sections. Test scores seldom are used as the sole basis of selection, so that in fact the test score is not the explicit selection variable but only an incidental selection variable. Unfortunately it is usually difficult in practice to quantify the actual explicit selection variable; hence it is difficult to apply the selection formula, and to treat the test score as an incidental selection variable. Treating an incidental selection variable as an explicit selection variable typically results in an undercorrection.

For the problem at hand, the exercise of some care in data gathering can be helpful in increasing the accuracy of the correction. First, it is necessary to isolate an explicit selection variable. This can often be done by arranging for the selection of all applicants whose test score is above some specified value. Applicants whose scores are below this value can also be selected as desirable, but for statistical calculations these applicants are not considered part of the selected group. This compromise arrangement should prove acceptable, administratively, in many situations.

Second, caution must be used to ensure that formulas are not applied when extreme selection has taken place. It is quite clear that the accuracy of the formulas decreases as the selection ratio increases. If tests must be compared in the presence of extreme selection on one variable, then other methods of comparison must be sought. Also, when dealing with problems involving very drastic selection, the formulation and methods of Chapters 17 through 20 may be more appropriate than methods of correlational analysis.

### Exercises

- 6.1. In the context of Theorem 6.2.1, prove that  $0 < (\sigma_x^2/\sigma_{\bar{X}}^2)(1 - \rho_{xT}^2) < 1$ .
- 6.2. Suppose we have two populations on which the same measurement is made. Further suppose that the observed score variances in the two populations are 100 and 90 and that the reliability of the measurement in the first population is 0.90. Assuming that the average standard error of measurement is the same for each group, determine the reliability of the measurement in the second population. For each population, compute the signal-to-noise ratio, and, for the second population, determine the factor by which the measurement must be lengthened so that its reliability in that population is equal to the original reliability in the first population.
- 6.3. The restriction of range correction for validity (Section 6.8) requires assumptions of linearity and homoscedasticity. Why does the restriction of range correction for reliability (Section 6.2) require only an assumption of homoscedasticity?
- 6.4. In Section 4.4, a composite reliability of 0.95 was reported for a *Test of English as a Foreign Language*. Can you now speculate as to why this unusually high value was obtained?
- 6.5. Show that as a corollary to Theorem 6.8.3 we may write

$$\left[ \frac{\sigma^2(Y)}{\sigma^2(Y^*)} - 1 \right] = \rho^2(X^*, Y^*) \left[ \frac{\sigma^2(X)}{\sigma^2(X^*)} - 1 \right].$$

- 6.6. As a complement to Theorem 6.8.3, show that if  $\sigma_Y^2$  is known rather than  $\sigma_X^2$ , we may write

$$\rho^2(X, Y) = \left\{ 1 - [1 - \rho^2(X^*, Y^*)] \frac{\sigma^2(Y^*)}{\sigma^2(Y)} \right\}$$

and

$$\sigma^2(X) = \frac{\sigma^2(X^*)[\sigma^2(Y) - \sigma^2(Y^*) - \rho^2(X^*, Y^*)\sigma^2(Y^*)]}{\rho^2(X^*, Y^*)\sigma^2(Y^*)}.$$

### References and Selected Readings

- BIRNBAUM, Z. W., On the effect of the cutting score when selection is performed against a dichotomized criterion. *Psychometrika*, 1950, **15**, 385–390.
- BIRNBAUM, Z. W., E. PAULSON, and F. C. ANDREWS, On the effect of selection performed on some coordinance of the multidimensional population. *Psychometrika*, 1950, **15**, 191–204.
- BOLDT, R. F., Study of linearity and homoscedasticity of test scores in the chance range. *Research Bulletin 66-43*. Princeton, N.J.: Educational Testing Service, 1966.
- COHEN, A. C., Restriction and selection in multinormal distributions. *Annals of Mathematical Statistics*, 1957, **28**, 731–741.
- COLEMAN, J. S., *Models of change and response uncertainty*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.

- CRONBACH, L. J., Test "reliability": its meaning and determination. *Psychometrika*, 1947, **12**, 1-16.
- CRONBACH, L. J., Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297-334.
- CRONBACH, L. J., and H. AZUMA, Internal-consistency reliability formulas applied to randomly sampled single-factor tests: an empirical comparison. *Educational and Psychological Measurement*, 1962, **22**, 645-665.
- CRONBACH, L. J., NAGESWARI RAJARATNAM, and GOLDINE GLESER, Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, **16**, 137-163.
- CRONBACH, L. J., and W. G. WARRINGTON, Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika*, 1951, **16**, 167-188.
- CURETON, E. E., The definition and estimation of test reliability. *Educational and Psychological Measurement*, 1958, **18**, 715-738.
- CURETON, E. E., Reliability and validity: basic assumptions and experimental designs. *Educational and Psychological Measurement*, 1965, **25**, 327-346.
- FRENCH, J. W., and W. B. MICHAEL, Standards for educational and psychological tests and manuals. Washington, D.C.: American Psychological Association, 1966.
- GARSIDE, R. F., The measurement of function fluctuation. *Psychometrika*, 1958, **23**, 75-84.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950, pp. 197-198. (a)
- GULLIKSEN, H., The reliability of speeded tests. *Psychometrika*, 1950, **15**, 259-269. (b)
- GUTTMAN, L., Reliability formulas for noncompleted or speeded tests. *Psychometrika*, 1955, **20**, 113-124.
- HOFFMAN, P. J., Test reliability and practice effects. *Psychometrika*, 1963, **28**, 273-288.
- HOWARD, K. I., Differentiation of individuals as a function of repeated testing. *Educational and Psychological Measurement*, 1964, **24**, 875-895.
- KELLEY, T. L., *Statistical methods*. New York: Macmillan, 1923, pp. 221-223.
- LAWLEY, D., A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh, Proceedings, Section A*, 1943-4, **62**, 28-30.
- LINN, R. L., Range restriction problems in the validation of a guidance test battery. *Research Bulletin 67-8*. Princeton, N.J.: Educational Testing Service, 1967.
- MAGNUSEN, D., *Introduction to test theory*. Reading, Mass.: Addison-Wesley, 1966.
- PETERS, C. C., and W. R. VAN VOORHIS, *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940, pp. 208-212.
- RYDBERG, S., *Bias in prediction*. Stockholm: Almqvist and Wiksell, 1963.
- SPEARMAN, C., Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 1907, **18**, 161-169.
- THORNDIKE, R. L., *Personnel selection*. New York: Wiley, 1949.
- VINSONHALER, J. F., and W. MEREDITH, A stochastic model for repeated testing. *Multivariate Behavioral Research*, 1966, **4**, 461-478.

# SOME ESTIMATES OF PARAMETERS OF THE CLASSICAL MODEL

## 7.1 Introduction

The earlier part of this book has dealt almost entirely with population parameters. Chapters 7 through 9 and 11 are primarily concerned with making inferences from samples.

If two or more parallel measurements are available from appropriately drawn samples of very large size, the formulas of Chapters 3 through 5 can be used to estimate the parameters of the classical model. If sample sizes are more moderate, or when the assumption of parallelism is discomforting, the more refined methods of estimation developed in this chapter are required.

To start with, in Section 7.2 we discuss the estimation of true scores. The remainder of Chapter 7 is concerned with the estimation of variances. We first assume that available repeated measurements on individuals are (at least approximately) parallel. We use this assumption in Section 7.3 to present a simple method for obtaining unbiased estimates of the variance of the (specific) errors of measurement. The practical use of estimated error variances is discussed briefly in Section 7.4. Section 7.6 shows how these same estimates of error variances can be obtained from a standard analysis of variance components. Estimates of true-score variance are also obtained from this analysis. A more general variance components analysis is introduced in Section 7.6, and the concept of generic error variance is introduced and contrasted with the specific error variance studied previously.

In Section 7.7, we consider the problem of estimating error variance when the parallelism assumption is unacceptable. The approach here is to obtain an unbiased estimate of a quantity that is an upper bound on the specific error variance. (Other methods for estimating specific error variance without parallel forms are treated at the end of Chapter 9 and in Chapter 10.)

This chapter covers only basic experimental design. Detailed analyses for separately estimating the effects of changes over time, changes in order of presentation of materials, and so forth are beyond the scope of this book. Some relevant references, however, are given at the ends of Sections 7.6 and 9.9.

## 7.2 Estimating True Score

By definition, an *unbiased estimator* of a parameter is an estimator whose expected value is the parameter estimated. Since  $\tau_a \equiv \mathcal{E}_k X_{ak}$ , it is clear that for a specified person  $a$ ,  $X_{ak}$  is an unbiased estimator of  $\tau_a$ .

We can easily see from the definition that a weighted sum of unbiased estimators is itself an unbiased estimator of a similarly weighted sum of parameters. Thus, denoting parameters by  $\theta$ , unbiased estimators by “hats” ( $\hat{\cdot}$ ), and weights by  $w$ , we may write

$$\mathcal{E} \sum_r w_r \hat{\theta}_r = \sum_r w_r \mathcal{E} \hat{\theta}_r = \sum_r w_r \theta_r. \quad (7.2.1)$$

We shall use this principle frequently.

From this principle it follows that a sample mean is an unbiased estimator of the corresponding population mean. Thus, in random sampling over people  $a$  and parallel measurements  $k$ , the sample average of  $N$  observed test scores

$$\bar{X} \equiv \frac{1}{N} \sum_a X_{ak},$$

is an unbiased estimator of  $\mu_T$ , the population mean of the true scores. This is seen from the equations

$$\begin{aligned} \mathcal{E}_k \frac{1}{N} \sum_{a=1}^N X_{ak} &= \frac{1}{N} \sum_a \mathcal{E}_k X_{ak} = \frac{1}{N} \sum_a \tau_a \equiv \bar{\tau}, \\ \mathcal{E}_a \bar{\tau} &= \frac{1}{N} \sum_a \mathcal{E}_a \tau_a = \frac{1}{N} \sum_a \mu_T = \mu_T. \end{aligned}$$

Although  $X_a$  may be an appropriate estimate of  $\tau_a$  when only one examinee has been tested, we may use a different estimate when certain information on the population of examinees is available. As shown in (3.7.2a), the linear mean squared error regression estimate of the true score of examinee  $a$  is not  $x_a$  but rather

$$R(T_a | x_a) = \mu_X + \rho_{XX'}(x_a - \mu_X). \quad (7.2.2)$$

Since  $\rho_{XX'}$  is positive, this estimate is larger than  $x_a$  when  $x_a < \mu_X$  and smaller than  $x_a$  when  $x_a > \mu_X$ .

The foregoing suggests that the observed score tends to underestimate (overestimate)  $\tau_a$  whenever  $x_a$  is below (above)  $\mu_X$ . If the reliability  $\rho_{XX'}$  is small, the difference between  $X_a$  and the regression estimate of (7.2.3) can be nearly as large as half the range of  $X$ . This is a serious difference from any practical point of view.

If a very large sample is available, the sample correlation  $r_{XX'}$  may be substituted for  $\rho_{XX'}$ , and  $\bar{x}$  for  $\mu_X$ , so that an approximate linear minimum mean squared error regression estimate of  $T_a$  is

$$\hat{R}(T_a | x_a) = r_{XX'} x_a + (1 - r_{XX'}) \bar{x}. \quad (7.2.3)$$

For example, if John's observed I.Q. is  $x_a = 180$ , and if  $\rho_{XX'} = 0.85$  and  $\mu_X = 100$  in a population of children to which John belongs, then by (7.2.2) the linear mean squared error regression estimate of his true I.Q., based on the information given, is  $(0.85)(180) + (0.15)(100) = 168$ . (If John is also a member of other populations, different estimates of his true I.Q. will, quite appropriately, be obtained, each estimate of course depending on the information utilized.)

Although the minimum mean squared error regression estimate of  $T_a$  is, in a certain sense, the best *linear* function of  $x_a$  for estimating  $T_a$ , it is not a very good estimator when the true regression of  $T$  on  $x$  is not approximately linear. Although approximate linearity is probably common in practice, there is reason not to expect strict linearity (see Section 22.9). In fact, if we were to find linearity in some particular population of examinees having a certain frequency distribution of true scores, then we would not be likely to find linearity in selected subpopulations having different distributions of true scores (for example, see Lord, 1959, Fig. 1).

Improvements on these linear estimates of  $T$  are possible when the data satisfy some stronger assumptions. A few such procedures are discussed in Sections 22.6 and 23.5. The method outlined in Section 23.5 is available for practical application when samples are sufficiently large. This method requires considerable computational effort, however.

### 7.3 An Unbiased Estimate of the Specific Error Variance from Parallel Measurements

If a random sample of  $N$  examinees is drawn from an infinite population (an infinitely large population of examinees is assumed in Chapters 7, 8, 9, and 10, except where otherwise specified), the expected value of the sample variance

$$s_y^2 \equiv \frac{1}{N} \sum_{a=1}^N (y_a - \bar{y})^2 \quad (7.3.1)$$

is

$$\mathcal{E}_a s_y^2 = \frac{N-1}{N} \sigma_Y^2, \quad (7.3.2)$$

where  $\sigma_Y^2 \equiv \mathcal{E}_a(Y_a - \mathcal{E}_a Y_a)^2$  is the corresponding population variance. The usual unbiased estimator of the population variance will be written  $\hat{\sigma}^2$ ; for example,

$$\hat{\sigma}_Y^2 \equiv \frac{N}{N-1} s_y^2 = \frac{1}{N-1} \sum_{a=1}^N (y_a - \bar{y})^2. \quad (7.3.3)$$

It should be noted that throughout this book the term *sample variance* refers to the second central moment of the observations, not to an unbiased estimate of the population variance. The symbol ( $\hat{\cdot}$ ) denotes a sample estimate, whether biased or unbiased.

The following two sections are concerned with parallel forms (replicate measurements) for a single test  $g$ . For the purposes of these two sections, the subscript  $g$  could simply be dropped from all symbols. We shall not do this, however, since it would confuse the relationship between the parameters and statistics studied in these sections and those studied in Section 7.7 and in Chapter 8, where we consider many *nearly* parallel tests  $g = 1, 2, \dots, n$  (there being no strictly parallel measurements available for any test).

As in Section 2.6, the error variance of examinee  $a$  is defined as the variance of the errors over parallel measurements on this examinee:

$$\sigma^2(E_{ga}) \equiv \sigma^2(E_{ga*}) \equiv \mathcal{E}_k(E_{gak} - \mathcal{E}_k E_{gak})^2 = \mathcal{E}_k E_{gak}^2, \quad (7.3.4)$$

where the subscript  $k$  denotes the  $k$ th replication or the  $k$ th parallel measurement. In Chapter 8 we shall see that it is possible to define more than one kind of error variance. We call the error variance (7.3.4) the *specific error variance for examinee  $a$  on measurement  $g$* , to distinguish it from others.

Suppose just two parallel measurements  $Y_{gak}$  and  $Y_{gal}$  are available for a single examinee. We shall show that an unbiased estimate of the examinee's specific error variance can be obtained from the difference

$$D_{ga} \equiv Y_{gal} - Y_{gak} \quad (7.3.5)$$

between two such parallel measurements. Since

$$Y_{gak'} = \tau_{ga} + E_{gak'}, \quad k' = k, l, \quad (7.3.6)$$

we have the important (if obvious) result that *the difference between two parallel measurements on the same examinee depends entirely on the errors of measurement*:

$$D_{ga} = \tau_{ga} + E_{gal} - (\tau_{ga} + E_{gak}) = E_{gal} - E_{gak}. \quad (7.3.7)$$

Since the correlation between errors on experimentally independent measurements is zero, the expected value of  $D_{ga}^2$  over all possible pairs of parallel measurements is

$$\begin{aligned} \mathcal{E} D_{ga}^2 &= \sigma^2(D_{ga}) = \sigma^2(E_{gal} - E_{gak}) = \sigma_l^2(E_{gal}) + \sigma_k^2(E_{gak}) \\ &= 2\sigma^2(E_{ga}), \end{aligned} \quad (7.3.8)$$

twice the specific error variance for examinee  $a$ .

Let  $d_{ga}$  denote the observed value of  $D_{ga}$ . Thus *the observable quantity  $\frac{1}{2} d_{ga}^2$  is an estimate, unbiased over parallel measurements, of the specific error variance for examinee  $a$* . This unbiased estimate is denoted by

$$\hat{\sigma}^2(E_{ga}) = \frac{1}{2} d_{ga}^2. \quad (7.3.9)$$

By (7.3.8),

$$\mathcal{E} \hat{\sigma}^2(E_{ga}) = \mathcal{E}_k \hat{\sigma}^2(E_{gak}) = \sigma^2(E_{ga}). \quad (7.3.10)$$

If two parallel measures on examinee  $a$  are 67 and 53, we can estimate that for examinee  $a$ , the standard deviation over parallel measurements of the specific errors is  $(67 - 53)/\sqrt{2}$ , or approximately ten score points.

Whenever a physical scientist wants to approximate a measurement on an object, he is usually able to take many parallel measurements without altering the object. If the psychometrician could do the same with examinees, he would have little trouble in estimating the error variance. Since the psychometrician's estimates are usually based on only two or at most only a few parallel measurements, however, his estimates of error variances for individual examinees are usually subject to excessively large sampling fluctuations. For this and other reasons, test theory often concerns itself with estimating  $\sigma^2(E_{g*})$ , the variance *over people* of the errors of measurement  $E_{ga}$ . As shown in (2.6.2), this variance is equal to the specific error variance  $\sigma^2(E_{ga})$  averaged over the population of examinees:

$$\sigma^2(E_{g*}) = \mathcal{E}_a \sigma^2(E_{ga}). \quad (7.3.11)$$

The quantity  $\sigma^2(E_{g*})$  is referred to as the *group specific error variance* or often simply as the *specific error variance*. It is the same quantity that was discussed in (2.7.3a) and that has been denoted in earlier chapters simply by  $\sigma_E^2$ . If all examinees had the same error variance, the group specific error variance would be appropriate for each examinee. Although such equality of error variances can seldom be assumed, the group specific error variance is often used as a rough approximation to the individual error variances of which it is the average. A better approximation can be obtained by breaking down the total group of examinees into subgroups of approximately equal ability, and then determining the group specific error variance separately for each subgroup.

An estimate  $\hat{\sigma}^2(E_{g*})$  for the group specific error variance is obtained by averaging the unbiased estimates for the error variances of individual examinees:

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{N} \sum_{a=1}^N \hat{\sigma}^2(E_{ga}) = \frac{1}{2N} \sum_{a=1}^N d_{ga}^2. \quad (7.3.12)$$

This estimate is unbiased with respect to random sampling of examinees, for by (7.3.12),

$$\mathcal{E}_a \hat{\sigma}^2(E_{g*}) = \frac{1}{N} \sum_{a=1}^N \mathcal{E}_a \hat{\sigma}^2(E_{gak}).$$

Since  $\mathcal{E}_a \hat{\sigma}^2(E_{ga})$  is a parameter, it must be the same for every replication (see Section 2.12) and hence it is not changed by writing  $\mathcal{E}_k \mathcal{E}_a \hat{\sigma}^2(E_{gak})$ . From (7.3.10) and (7.3.11), we therefore have

$$\begin{aligned} \mathcal{E}_a \hat{\sigma}^2(E_{g*}) &= \frac{1}{N} \sum_{a=1}^N \mathcal{E}_a \mathcal{E}_k \hat{\sigma}^2(E_{gak}) = \frac{1}{N} \sum_{a=1}^N \mathcal{E}_a \sigma^2(E_{ga}) \\ &= \frac{1}{N} \sum_{a=1}^N \sigma^2(E_{g*}) = \sigma^2(E_{g*}). \end{aligned}$$

**Table 7.3.1**

Computations of estimates of error variance, true-score variance, and reliability

<b>A. Data</b>	1	2	3	4	5	6	7	8	9	10	Total
Person(s) $a$	$y_{ga1}$	125	119	109	104	101	98	97	94	90	81 1018
	$y_{ga2}$	120	122	107	108	98	106	96	99	93	87 1036

**B. Preliminary computations.**  $N = 10$ ,  $r = 2$ :

$$y_{ga+} \equiv \sum_{k=1}^r y_{gak} = 245 + 241 + 216 + 212 + 199 + 204 + 193 + 193 + 183 + 168 = 2054$$

$$d_{ga} = y_{ga2} - y_{ga1} = 5 - 3 = 2$$

$$d_{ga} = y_{ga2} - y_{ga1} = 5 - 3 = 2$$

$$\sum_{a=1}^N y_{ga1} = 1018 \quad \sum_{a=1}^N y_{ga2} = 1036$$

$$y_{g++} \equiv \sum_{a=1}^N y_{ga+} = \sum_{a=1}^N \sum_{k=1}^r y_{gak} = 2054$$

$$\sum_{a=1}^N y_{ga1}^2 = 105,194 \quad \sum_{a=1}^N y_{ga2}^2 = 108,472$$

$$\sum_{a=1}^N \sum_{k=1}^r y_{gak}^2 = 105,194 + 108,472 = 213,666$$

$$\sum_{a=1}^N y_{ga+}^2 = 427,134 \quad \sum_{a=1}^N d_{ga}^2 = 198 \quad \sum_{a=1}^N y_{ga1} y_{ga2} = 106,734$$

**C. Estimated group specific error variance**

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{2N} \sum_{a=1}^N d_{ga}^2 = \frac{1}{2(10)} (198) = 9.9, \quad (7.3.12)$$

or

$$\begin{aligned} \hat{\sigma}^2(E_{g*}) &= \frac{1}{N(r-1)} \sum_{a=1}^N \sum_{k=1}^r (y_{gak} - y_{ga+})^2 \\ &= \frac{1}{N(r-1)} \left[ \sum_{a=1}^N \sum_{k=1}^r y_{gak}^2 - \frac{1}{r} \sum_{a=1}^N \left( \sum_{k=1}^r y_{gak} \right)^2 \right] \\ &= \frac{1}{N(r-1)} \left( \sum_{a=1}^N \sum_{k=1}^r y_{gak}^2 - \frac{1}{r} \sum_{a=1}^N y_{ga+}^2 \right) \\ &= \frac{1}{10(2-1)} \left( 213,666 - \frac{427,134}{2} \right) = 9.9. \end{aligned} \quad (7.3.17)$$

**Table 7.3.1 (cont.)**

D. The following computations will be used in conjunction with Sections 7.5, 7.7, and 9.2:

1) The estimated specific true-score variance is obtained from

$$\begin{aligned}s^2(y_{g*}) &= \frac{1}{N} \sum_{a=1}^N (y_{ga*} - y_{g*..})^2 \\&= \frac{1}{Nr^2} \left[ \sum_{a=1}^N \left( \sum_{k=1}^r y_{gak} \right)^2 - \frac{1}{N} \left( \sum_{a=1}^N \sum_{k=1}^r y_{gak} \right)^2 \right] = \frac{1}{Nr^2} \left( \sum_{a=1}^N y_{ga+}^2 - \frac{y_{g++}^2}{N} \right) \\&= \frac{1}{10(2)^2} \left( 427,134 - \frac{(2054)^2}{10} \right) = 131.06,\end{aligned}\quad (7.5.3)$$

and

$$\hat{\sigma}^2(T_{g*}) = \frac{N}{N-1} s^2(y_{g*..}) - \frac{1}{r} \hat{\sigma}^2(E_{g*}) = \frac{1}{9}(131.06) - \frac{1}{2}(9.9) = 140.67. \quad (7.5.2)$$


---

2) The specific reliability may be estimated from

$$\begin{aligned}\hat{\rho}_{YT}^2 &= \frac{\hat{\sigma}^2(T_{g*})}{\hat{\sigma}^2(T_{g*}) + \hat{\sigma}^2(E_{g*})} \\&= \frac{140.67}{140.67 + 9.9} = 0.9342.\end{aligned}\quad (9.2.4)$$

We prefer this to use of the

$$\begin{aligned}\text{among-persons mean square} &= \frac{1}{r} \frac{N}{N-1} \sum_{k=1}^r s^2(y_{g*k}) \\&= \frac{1}{r} \frac{1}{N(N-1)} \sum_{k=1}^r \left[ N \sum_a y_{gak}^2 - \left( \sum_a y_{gak} \right)^2 \right] \\&= \frac{1}{180}(15,616 + 11,424) = 150.22. \\ \frac{\hat{\sigma}^2(T_{g*})}{\text{among-persons mean square}} &= \frac{140.67}{150.22} = 0.9364. \text{ (See Section 9.2.)}\end{aligned}\quad (9.2.2)$$


---

3) The squared correlation between parallel measurements is

$$\begin{aligned}r^2(y_{g*1}, y_{g*2}) &= \frac{\left( N \sum_{a=1}^N y_{ga1} y_{ga2} - \sum_{a=1}^N y_{ga1} \sum_{a=1}^N y_{ga2} \right)^2}{\left[ N \sum_{a=1}^N y_{ga1}^2 - \left( \sum_{a=1}^N y_{ga1} \right)^2 \right] \left[ N \sum_{a=1}^N y_{ga2}^2 - \left( \sum_{a=1}^N y_{ga2} \right)^2 \right]} \\&= \frac{[10(106,734) - (1018)(1036)]^2}{[10(105,194) - (1018)^2][10(108,472) - (1036)^2]} \\&= \frac{(12,692)^2}{(15,616)(11,424)} = \frac{161,086,864}{178,397,184} = 0.9030.\end{aligned}$$


---

If there are just two parallel measurements on each examinee, (7.3.12) gives an estimate, unbiased with respect to random sampling of examinees, for the group specific error variance (the variance over individuals of the errors of measurement).

If there are  $r \geq 2$  parallel measurements, an unbiased estimate of  $\sigma^2(E_{ga})$ , the individual's specific error variance, is given by

$$\hat{\sigma}^2(E_{ga}) \equiv \frac{r}{r-1} s^2(y_{ga*}), \quad (7.3.13)$$

where  $s^2(y_{ga*})$  is the sample variance of the scores of examinee  $a$  over parallel measurements:

$$s^2(y_{ga*}) \equiv \frac{1}{r} \sum_{k=1}^r (y_{gak} - y_{ga*})^2. \quad (7.3.14)$$

The dot subscript here and elsewhere indicates the sample average taken over the missing subscript. Thus

$$y_{ga*} \equiv \frac{1}{r} \sum_{k=1}^r y_{gak}. \quad (7.3.15)$$

That the right-hand side of (7.3.13) is an unbiased estimate of  $\sigma^2(Y_{ga*})$  may be seen by noting that for a single examinee,  $Y_{ga}$  and  $E_{ga}$  will have equal variances, since the two variables differ by a constant amount  $\tau_{ga}$ . When  $r = 2$ , (7.3.13) reduces to (7.3.9).

As in (7.3.12), we may write an estimate of the group specific error variance, unbiased in random sampling of examinees:

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{N} \sum_{a=1}^N \hat{\sigma}^2(E_{ga}). \quad (7.3.16)$$

If  $r$  parallel measurements are available on each examinee, (7.3.16) gives an estimate, unbiased with respect to random sampling of examinees, for the group specific error variance (the variance over individuals of the errors of measurement). Equation (7.3.16) can be written explicitly as

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{N(r-1)} \sum_{a=1}^N \sum_{k=1}^r (y_{gak} - y_{ga*})^2. \quad (7.3.17)$$

The estimate  $\hat{\sigma}^2(E_{g*})$  of (7.3.17) is the estimate that arises in a standard analysis of variance components, as shown in Table 7.5.1.

In Table 7.3.1, we present some hypothetical data and computations of estimates of error variance, true-score variance, and reliability. (The reader should refer to Section 7.5 in connection with true-score variance, and Section 9.2 in connection with reliability.) In Exercise 7.2, we provide a second set of data and ask the reader to perform similar computations.

#### 7.4 The Use of Estimated Error Variances

It seems obvious that if we have measurements containing errors, we will wish to know something about the size of the errors. As an index of magnitude, we commonly use the square root of the estimated group error variance, which we call the *estimated group standard error of measurement*. Our thinking here is that an examinee's true score "probably" lies somewhere within a band extending, say, two standard errors on each side of his actual observed score.

For example, suppose that a student obtains a score of  $y_{1a} = 91$  on a pre-test given at the beginning of the school year and a score of  $y_{2a} = 98$  on a parallel test form given at the end of the year. (The reader should note that the measurements thus obtained are not parallel. Parallel forms administered at the same time and under the same conditions in theory yield parallel measurements; since there is a difference in time in this example, we have parallel forms but not parallel measurements). Further, let us say that the group specific error variance is reliably estimated from (7.3.12) or (7.3.17) to be 9.9. For illustrative purposes, we have chosen this value to be the same as the value obtained in the computation that follows Table 7.3.1. Now suppose that earlier studies have shown that

- 1) the errors of measurement for a single student are approximately normally distributed,
- 2) the individual error variances for students in the score range of 90–100 are approximately the same as the group error variance, and
- 3) there is no appreciable practice effect when these two parallel forms are administered several months apart.

Under these conditions, the difference  $Y_{2a} - Y_{1a}$  is normally distributed about the corresponding true-score difference  $\tau_{2a} - \tau_{1a}$  with a variance of

$$\sigma^2(Y_{2a} - Y_{1a}) = \sigma^2(E_{2a} - E_{1a}) = 2\sigma_E^2 = 19.8.$$

Thus a 95% confidence interval for  $\tau_2 - \tau_1$  is

$$7 - 1.96\sqrt{19.8} \leq \tau_2 - \tau_1 \leq 7 + 1.96\sqrt{19.8},$$

or

$$-1.72 \leq \tau_2 - \tau_1 \leq 15.72,$$

which leaves open the possibility that there may not have been any true gain at all.

Similarly, if John gets a score of  $y_{oa} = 91$  and Mary gets a score of  $y_{ob} = 98$  on the same testing, the same analysis may be used to make a corresponding inference about the true difference in the examinees' ability. If many such inferences are made over a period of time about many students chosen without

reference to their observed test scores, one can have confidence that most of these inferences are correct. On the other hand, one *cannot* have confidence in the correctness of such an inference for any single, nonrandomly chosen individual or pair of individuals. In particular, it should be understood that confidence intervals obtained in this way are not valid if we become interested in John and Mary as a result of seeing their test scores.

### 7.5 Specific True-Score Variance Estimated from an Analysis of Variance Components

An estimate of the true-score variance, of interest for its own sake, will be needed for estimating reliability coefficients in Chapter 9. In this section we develop the components analysis estimate of the true-score variance, making the assumption that replicate measurements are available. In this section and in other discussions of components analysis, we shall frequently use the term *replications* instead of *parallel measurements* since the usage and implications of the former term are well established in this context. In theory, replicate measurements are those obtained by retesting either with the same form or with parallel forms. In actual testing neither method is likely to yield strictly parallel measurements because of practice effects. Note that in this section we are considering only strictly parallel measurements, *not* the possibly nonparallel measurements obtained from parallel forms.

When  $r \geq 2$  replicate measurements are available on a single test  $g$ , the results of the usual components analysis would appear as in Table 7.5.1. Note that there is no entry for an *Among replications* sum of squares with  $r - 1$  degrees of freedom, nor for an *Interaction* sum of squares. (Differences among replications are, by definition of replication, due to the same chance fluctuations that give rise to the "Replications" sum of squares shown.)

**Table 7.5.1**  
Components analysis for  
replications of a single test\*

Source	Sum of squares	Degrees of freedom	Expected mean square
Among persons	$r \sum_{a=1}^N (y_{ga.} - y_{g..})^2$	$N - 1$	$\sigma^2(E_{ga}) + r\sigma^2(T_{ga})$
Replications	$\sum_a \sum_{k=1}^r (y_{gak} - y_{ga.})^2$	$N(r - 1)$	$\sigma^2(E_{gak})$
Total	$\sum_a \sum_k (y_{gak} - y_{g..})^2$	$Nr - 1$	

\* The model is  $y_{gak} \equiv \mu + (T_{ga} - \mu) + E_{gak}$ .

It is clear from the last column that the components analysis unbiased estimate of  $\sigma^2(T_{g*})$  is

$$\hat{\sigma}^2(T_{g*}) = \frac{1}{r} (\text{among-persons mean square} - \text{replications mean square}). \quad (7.5.1)$$

By virtue of (7.3.17), this becomes

$$\hat{\sigma}^2(T_{g*}) = \frac{N}{N-1} s^2(y_{g*..}) - \frac{1}{r} \hat{\sigma}^2(E_{g*}), \quad (7.5.2)$$

where

$$s^2(y_{g*..}) \equiv \frac{1}{N} \sum_{a=1}^N (y_{ga..} - y_{g..})^2. \quad (7.5.3)$$

Equation (7.5.2) gives an estimate for the specific true-score variance  $\sigma^2(T_{g*})$  that is unbiased with respect to random sampling of examinees. This is the same true-score variance discussed in Chapters 2 through 6.

For  $r = 2$ , (7.5.2) does not assume a particularly simple form if rewritten in terms of the correlation between replicate measurements. However, by (3.3.5c),

$$\sigma^2(T_{g*}) = \sigma(Y_{g*1}, Y_{g*2}). \quad (7.5.4)$$

Now

$$\frac{N}{N-1} s(y_{g*1}, y_{g*2}) \equiv \frac{1}{N-1} \sum_{a=1}^N (y_{ga1} - y_{g..1})(y_{ga2} - y_{g..2}) \quad (7.5.5)$$

is an unbiased estimator of the population covariance  $\sigma(Y_{g*1}, Y_{g*2})$ . Consequently  $N/(N-1)$  times the sample covariance between replicate measurements is an unbiased estimate of the specific true-score variance  $\sigma^2(T_{g*})$ .

When  $r = 2$ , (7.5.2) and (7.5.5) illustrate the fact that there are many unbiased estimates for a single population parameter. It may be instructive to suggest the nature of the difference between the two estimates (see also Exercise 7.6). Note first of all that  $s(y_{g*1}, y_{g*2})$  would not be changed if a constant were added to  $y_{ga2}$  for each examinee; on the other hand,  $\hat{\sigma}^2(E_{g*})$  would definitely be changed, as is clear from (7.3.12), and consequently so would  $\hat{\sigma}^2(T_{g*})$  in (7.5.2). The fact is that  $s(y_{g*1}, y_{g*2})$  does not depend on, and hence does not use, the information available in the mean difference

$$d_{g*} \equiv \frac{1}{N} \sum_a d_{ga}. \quad (7.5.6)$$

If the reader applies (7.5.2) and (7.5.5) to the data in Table 7.3.1, he will find that

$$\hat{\sigma}^2(T_{g*}) = 140.67, \quad \frac{N}{N-1} s(y_{g*1}, y_{g*2}) = 141.02.$$

It should be noted that  $\hat{\sigma}^2(T_{g*})$  and also  $s(y_{g*1}, y_{g*2})$  may happen to be negative because of sampling fluctuations. If  $\hat{\sigma}^2(T_{g*}) > 0$ , then  $\hat{\sigma}^2(T_{g*})$  and  $s(y_{g*1}, y_{g*2})$  are both unbiased estimates of a positive quantity. Since the sampling variance of either estimate approaches zero as either the number of examinees or the number of parallel test forms becomes large, the probability that either estimate is negative also approaches zero. A discussion of negative estimates of variance is given by Thompson (1962), by Hill (1965), and by Tiao and Tan (1965).

Here we shall briefly illustrate a use of an estimated true-score variance. Suppose the mean squares computed from Table 7.5.1 for 100 persons and 3 replications are 196 (among persons) and 215 (replications). By (7.5.1),  $\hat{\sigma}^2(T_{g*})$  is slightly negative. The conclusion to be drawn is that all observed-score differences among persons may be entirely due to sampling fluctuations. In other words, the true-score variance may actually be nearly zero, indicating either that the examinees differ only very little on the trait measured or, possibly, that the test as scored is capable of providing only very unreliable systematic measurement of any single trait.

Estimated true-score variances are most commonly used to obtain estimated reliability coefficients. This application is the subject of Chapter 9.

## 7.6 A General Formulation of the Estimation Problem as an Analysis of Variance Components

An unbiased estimate of error variance also can be obtained from Table 7.5.1. From the second row of the table, we see that

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{N(r-1)} \sum_a \sum_{k=1}^r (y_{gak} - y_{ga*})^2$$

is an unbiased estimate of the specific error variance. This is the same estimate that was given as (7.3.17).

Problems of estimating error variance and true-score variance are naturally and conveniently met by standard procedures for estimating variance components in analysis of variance. This approach is particularly helpful for comparing different variance estimates obtained from different sets of assumptions. For present purposes, components analysis constitutes a ready-made, detailed scheme for rigorous statistical inference. At the end of this section, we shall briefly discuss the extent to which optimal properties can be claimed for the estimates so obtained.

Table 7.6.1 displays a very general analysis of variance components for  $n$  different tests administered to  $N$  examinees with  $r$  replications. These  $n$  tests need not all measure exactly the same psychological trait. If  $n = 1$ , this table can be reduced to Table 7.5.1.

Strictly replicate measurements are rarely obtained in actual testing because of practice effects, chronological-order effects, and so forth. In principle, such effects can be dealt with by further complicating the variance components

**Table 7.6.1**  
Analysis of variance components for replicated test data

Source	Sum of squares	Degrees of freedom	Expected mean square
Among persons	$rn \sum_{a=1}^N (y_{a..} - y_{...})^2$	$N - 1$	$\sigma^2(E_{..}) + r \left(1 - \frac{n}{G}\right) \sigma_\alpha^2 + rn\sigma_\xi^2$
Among tests	$rN \sum_{g=1}^n (y_{g..} - y_{...})^2$	$n - 1$	$\sigma^2(E_{..}) + r\sigma_\alpha^2 + rN\sigma_\pi^2$
Interaction	$r \sum_g \sum_a (y_{ga..} - y_{g..} - y_{a..} + y_{...})^2$	$(N - 1)(n - 1)$	$\sigma^2(E_{..}) + r\sigma_\alpha^2$
Replications	$\sum_g \sum_a \sum_{k=1}^r (y_{gak} - y_{ga..})^2$	$nN(r - 1)$	$\sigma^2(E_{..})$
Total	$\sum_g \sum_a \sum_k (y_{gak} - y_{...})^2$	$nNr - 1$	

model, but this is something that we are anxious to avoid at this point. A further brief comment is given at the end of this section. For the present, we deal only with the idealized situation represented by Table 7.6.1. We outline the terminology, the notation, and the mathematical relationships that are necessary to make use of the table. General discussions of the development of this kind of variance component table may be found in Graybill (1961, Chapters 16, 18), Scheffé (1959, Chapters 7, 8), and Winer (1962, Chapter 5).

The last column of the table shows the expected mean square, i.e., the expected value of each sum of squares, under random and independent sampling of tests and people and replications, divided by the corresponding degrees of freedom. The reader who wishes may derive these expected values by familiar techniques that he has already employed in Chapters 4 through 7. The work is tedious but straightforward.

In an infinite population of examinees (assumed throughout), every replication has the same population parameters (see Section 2.12). Consequently the expected mean squares given in the table are valid for each replication taken singly.

There are no assumptions that any of the "effects" underlying Table 7.6.1 are normally distributed (see, for example, Cornfield and Tukey, 1956). In the absence of normality assumptions, significance tests of point hypotheses based on the  $F$ -distribution either cannot be made or require special justification. However, such significance tests of point hypotheses are of little value in reliability theory.

The sampling errors of the variance estimates discussed in this chapter and the next can be obtained by the methods of Chapter 11 without any assumption of normal distributions. Except in special cases, these sampling variances have complicated formulas that will not be given here; the reader can find some of these formulas in Hooke (1956) and Dayhoff (1966). Of course, small-sample

frequency distributions, and hence small-sample confidence intervals, for the variance estimates are unknown in the absence of normality or other assumptions about the distribution of the data.

In this and subsequent chapters, the custom of distinguishing a random variable from its observed value by the use of a capital letter will be discarded in the case of Greek letters having capitals likely to be confused with Roman letters. For present purposes, we shall assume that people and replications are each sampled from infinite populations, and that the  $n$  tests are drawn at random from a (possibly hypothetical) population of  $G$  tests, where either  $1 \leq n < G \leq \infty$  or  $1 \leq n \leq G < \infty$ . (We shall discuss the assumption of random sampling in relation to tests in some detail at the beginning of Chapter 11.) Note that this formulation includes the simplest case where  $n = G$ , that is, where the tests are not sampled at all but are considered as fixed. If  $n < G \leq \infty$ , then both persons and tests are taken at random, and the components analysis is based on a random-components model (finite or infinite). If  $n = G < \infty$ , then the persons are taken at random but the tests are fixed, and the components analysis is based on a mixed model. Cornfield and Tukey (1956) have formulated all these different possibilities in one table (or see Winer, 1962, Chapter 5).

Each of the variances in Table 7.6.1 has an important use in test theory, which we shall discuss in detail in Chapters 8 through 10. The symbol  $\sigma_{\zeta}^2$  refers to the variance over examinees of the *generic true score* defined by

$$\zeta_a \equiv \mathcal{E}_g \mathcal{E}_k Y_{gak} = \mathcal{E}_g T_{ga}. \quad (7.6.1)$$

The generic true score is a random variable over the population of examinees. It represents the standing of the examinee with respect to the average of the traits measured by all the  $G$  test forms. It is discussed in detail in Section 8.1. The quantity  $\zeta_a - \mu$ , where

$$\mu \equiv \mu_r \equiv \mathcal{E}_a \zeta_a, \quad (7.6.2)$$

is called the *persons effect*.

In the table, the symbol  $\sigma_{\pi}^2$  refers to the variance over tests of the *test difficulty*:

$$\begin{aligned} \pi_g &\equiv \mathcal{E}_a \mathcal{E}_k Y_{gak} = \mathcal{E}_a T_{ga} \\ &\equiv \mathcal{E}_k \mathcal{E}_a Y_{gak} = \mathcal{E}_a Y_{ga}. \end{aligned} \quad (7.6.3)$$

[The last equals sign is justified by the fact that  $\mathcal{E}_a Y_{gak}$  is a parameter that must be the same for every  $k$  (see Section 2.12)]. The *difficulty*  $\pi_g$  of test form  $g$  reflects the average level of examinee performance on this form. Note that by an unfortunate quirk of nomenclature (already well established), the more difficult a test is for the examinees, the lower the value of the difficulty index  $\pi_g$ . Each  $\pi_g$  may be considered a fixed constant for the  $G = n$  tests at hand; or  $\pi_g$  may be considered as a random variable over the population of  $G \leq \infty$  tests. The quantity  $\pi_g - \mu$  is called the *test-form effect*, or simply the *test effect*.

The symbol

$$\alpha_{ga} \equiv (T_{ga} - \mu) - (\zeta_a - \mu) - (\pi_g - \mu) \equiv T_{ga} - \zeta_a - \pi_g + \mu \quad (7.6.4)$$

denotes the *interaction effect* for person  $a$  on test  $g$  or, more simply, the *interaction*. If all test measurements  $g = 1, 2, \dots, G$  were strictly parallel, then we would have  $\pi_g = \mu$ ,  $\zeta_a = T_{ga}$ , and  $\alpha_{ga} = 0$ . If the interaction effects are large, this is a clear indication that the measurements are not parallel. For fixed  $g$ , the interaction may be considered as a random variable over the population of examinees; for fixed  $a$ , the interaction may be considered as a random variable over a population of tests.

We can see from (7.6.1), (7.6.2), (7.6.3), and (7.6.4) that the following expression for  $Y_{gak}$  is a mathematical identity:

$$Y_{gak} \equiv \mu + (\zeta_a - \mu) + (\pi_g - \mu) + \alpha_{ga} + E_{gak}, \quad (7.6.5)$$

$E_{gak}$  being the specific error of measurement defined by (7.3.6). *Equation (7.6.5) is called the linear model underlying Table 7.6.1.* It points out that the various “effects” have been defined in such a way that the observed score  $Y_{gak}$  may be thought of as a simple sum of these effects, plus a constant ( $\mu$ ) and an error of measurement ( $E_{gak}$ ). This further decomposition of the random variable  $Y_{gak}$  is a direct generalization of the method used in Chapter 2 to construct true and error random variables. The present decomposition allows us to construct generic true score, test difficulty, and interaction random variables that are uncorrelated both with each other and with a residual variable called an error.

The only symbol in Table 7.6.1 still to be discussed is  $\sigma^2(E_{**})$ . According to the last row in the body of the table, an unbiased estimate of  $\sigma^2(E_{**})$  is the mean square for replications:

$$\hat{\sigma}^2(E_{**}) = \frac{1}{nN(r-1)} \sum_g \sum_a \sum_{k=1}^r (y_{gak} - y_{ga*})^2. \quad (7.6.6)$$

From (7.3.17), we see that

$$\hat{\sigma}^2(E_{**}) = \frac{1}{n} \sum_{g=1}^n \hat{\sigma}^2(E_{g*}). \quad (7.6.7)$$

*The replication mean square is the average over all  $n$  tests of  $\hat{\sigma}^2(E_{g*})$ , the unbiased estimate (7.3.16) for the group specific error variance. If  $n = 1$ , the two estimates are identical.*

By taking the expectation in (7.6.7) over people and tests, we find that

$$\sigma^2(E_{**}) = \mathcal{E}_g \sigma^2(E_{g*}). \quad (7.6.8)$$

Thus  $\sigma^2(E_{**})$  is the expectation over all  $G$  tests of the group specific error variance  $\sigma^2(E_{g*})$  defined by (7.3.11).

The complete analysis shown in Table 7.6.1 is needed for theoretical rather than for computational purposes. If one has replicate measurements for test  $g$ , he does not usually wish to average the sums of squares for test  $g$  with those for other tests. The theoretical reason for considering Table 7.6.1 will become apparent in the next section.

Variance estimators obtained in components analysis are unbiased, but other optimum properties are not easy to prove without restrictive assumptions about the frequency distribution of the variables involved ( $Y$ ,  $\xi$ ,  $\pi$ ,  $\alpha$ , and  $E$ ). For the random effects model, a theorem by Graybill and Hultquist (1961, Section 7) shows that under certain assumptions the estimators obtained from components analyses such as those discussed in this chapter are *minimum variance unbiased quadratic estimators*. Their theorem requires that the “effects”  $\alpha_{ga}$ ,  $\xi_a - \mu$ , and  $\pi_g - \mu$  ( $g = 1, 2, \dots, n$ ;  $a = 1, 2, \dots, N$ ),

- 1) be independently distributed, and
- 2) have finite second, third, and fourth moments, and that
- 3) these moments be the same for all  $a$  and  $g$ .

Of these assumptions, both (1) and (3) are usually violated by mental test data. It is not known whether the Graybill-Hultquist theorem can be generalized to hold without these assumptions. For the present, however, components analysis estimators are recommended here both for theoretical and for careful practical work.

Again we point out that more complicated experimental design and analysis are required if there is practice effect or other chronological-order effect. Stanley (1955, 1962, and personal communication) suggests the following design to deal with chronological-order effects. Suppose there are  $r = 3$  forms of one test, called  $A$ ,  $B$ , and  $C$ . There are then six possible sequences of forms:  $ABC$ ,  $ACB$ ,  $BAC$ ,  $BCA$ ,  $CAB$ ,  $CBA$ . There are to be three separate test administrations: first, second, and third. Each examinee is to take all three forms, administered in one of the six possible sequences. Examinees are to be assigned to sequences at random. The main effects are sequence ( $ABC$ ,  $ACB$ ,  $\dots$ ,  $CBA$ ), test administration (first, second, third), test form ( $A$ ,  $B$ ,  $C$ ), and examinee nested within sequence. This experimental design deals with order effects directly, avoiding the unrealistic assumption that they do not exist. Space limitations do not permit a fuller discussion here; the reader is referred to Stanley's papers, to the references given there and at the end of Section 9.9 and particularly to Stanley's chapter, “Reliability”, to appear in Thorndike (in preparation).

## 7.7 An Estimate of an Upper Bound on the Specific Error Variance from Measurements That Are Not Strictly Parallel

In practice, one usually does not expect two supposedly parallel test forms to be of exactly equal difficulty; one really does not expect them to be strictly parallel. What shall be done in this situation? What is the effect of substituting

nearly parallel measures—we shall call them *nominally parallel* measures—in formulas that call for strictly parallel measures? These questions will be considered here and in Chapter 8. Since replications are not available, we shall drop the third subscript from all formulas.

When strictly parallel measurements are unavailable to the psychometrician and  $r$  is consequently equal to one, the expected mean square for replications of (7.6.8), namely,

$$\sigma^2(E_{**}) = \mathcal{E}_g \sigma^2(E_{g*}),$$

cannot be estimated from (7.6.6) because of the  $r - 1$  in the denominator. There simply is no information available for producing a consistent estimate of  $\sigma^2(E_{**})$ .

We can see from the next-to-last row of Table 7.6.1, however, that the interaction mean square ( $\overline{MSI}$ ) is an unbiased estimate, over random sampling of people and tests, for  $\sigma^2(E_{**}) + r\sigma_\alpha^2$ , which is a quantity that can never be less than  $\sigma^2(E_{**})$ . Thus *the interaction mean square is an unbiased estimate of an upper bound to the group specific error variance averaged over all G tests*. The bound is attained when  $\sigma_\alpha^2 = 0$ .

If the test forms are sufficiently similar so that  $\sigma_\alpha^2$  is sufficiently small, the interaction mean square provides an acceptable estimate of the specific error variance of any one of the forms. The important thing about this result is that  $\overline{MSI}$  can be computed when  $r = 1$ . This estimate of the group specific error variance is denoted by

$$\tilde{\sigma}_e^2 \equiv \overline{MSI} \equiv \frac{1}{(n - 1)(N - 1)} \sum_{g=1}^n \sum_{a=1}^N (y_{ga} - y_{g*} - y_{*a} + y_{..})^2, \quad (7.7.1)$$

the third subscript being omitted since there is no replication. As already noted,

$$\mathcal{E}_g \mathcal{E}_a \tilde{\sigma}_e^2 = \sigma^2(E_{**}) + \sigma_\alpha^2. \quad (7.7.2)$$

*Equation (7.7.1) provides an unbiased estimate, over random sampling of people and tests, of an upper bound to the group specific error variance averaged over all G tests.* Since it is a sample statistic,  $\tilde{\sigma}_e^2$  is not itself an upper bound; for convenience we shall sometimes refer to it as the *generous estimate* of the specific error variance.

Better computing formulas for  $\tilde{\sigma}_e^2$  may be derived algebraically. It is instructive to obtain one such formula as follows: If Table 7.6.1 is rewritten with  $r = 1$ , it becomes Table 7.7.1. We may obtain an alternative analysis, equally familiar in form, by appropriately combining the first and third rows of Table 7.7.1; this analysis appears as Table 7.7.2. The row labeled "Among tests" and also the row labeled "Total" are identical in these two tables. Since the sums of squares in the body of each table are known to add up to the total sum of squares, it follows that the within-tests sum of squares in Table 7.7.2 is equal to the among-persons sum of squares plus the interaction sum of squares in

**Table 7.7.1** Analysis of variance components for nonreplicated test data

Source	Sum of squares	Degrees of freedom	Expected mean square
Among persons	$n \sum_{a=1}^N (y_{\cdot a} - y_{..})^2 \equiv nNs^2(y_{..})$	$N - 1$	$\sigma^2(E_{**}) + \left(1 - \frac{n}{G}\right) \sigma_a^2 + n\sigma_i^2$
Among tests	$N \sum_{g=1}^n (y_{g\cdot} - y_{..})^2 \equiv nNs^2(y_{..})$	$n - 1$	$\sigma^2(E_{**}) + \sigma_a^2 + N\sigma_\pi^2$
Interaction	$\sum_g \sum_a (y_{ga} - y_{g\cdot} - y_{\cdot a} + y_{..})^2$ $\equiv (N - 1)(n - 1)\tilde{\sigma}_e^2$	$(N - 1)(n - 1)$	$\sigma^2(E_{**}) + \sigma_a^2$
Total	$\sum_g \sum_a (y_{ga} - y_{..})^2 \equiv nNs^2(y_{..})$	$nN - 1$	

**Table 7.7.2** Alternative analysis for nonreplicated data

Source	Sum of squares	Degrees of freedom	Expected mean square
Among tests	$N \sum_{g=1}^n (y_{g\cdot} - y_{..})^2 \equiv nNs^2(y_{..})$	$n - 1$	$\sigma^2(E_{**}) + \sigma_a^2 + N\sigma_\pi^2$
Within tests	$\sum_g \sum_{a=1}^N (y_{ga} - y_{g\cdot})^2 \equiv N \sum_g s^2(y_{g\cdot})$	$n(N - 1)$	$\sigma^2(E_{**}) + \sigma_a^2 + \sigma_i^2$
Total	$\sum_{g=1}^n \sum_{a=1}^N (y_{ga} - y_{..})^2 \equiv nNs^2(y_{..})$	$nN - 1$	

Table 7.7.1. This fact provides an alternative computing formula for  $\tilde{\sigma}_e^2$ ,

$$\tilde{\sigma}_e^2 = \frac{N}{(n-1)(N-1)} \left[ \sum_{g=1}^n s^2(y_{g*}) - ns^2(y_{..}) \right], \quad (7.7.3)$$

which the reader may derive algebraically. In this formula,

$$s^2(y_{g*}) \equiv \frac{1}{N} \sum_{a=1}^N (y_{ga} - y_{g*})^2, \quad (7.7.4)$$

and  $nNs^2(y_{..})$  is the among-persons sum of squares defined algebraically in the top row of Table 7.7.1.

When there are just two tests ( $n = 2$ ), we can write a simple and instructive formula for the generous estimate. Let

$$d_a \equiv y_{2a} - y_{1a}; \quad (7.7.5)$$

then

$$y_{ga} - y_{..} \equiv \frac{1}{2}(-1)^g d_a, \quad g = 1, 2, \quad (7.7.6)$$

$$y_{g*} - y_{..} = \frac{1}{2N} (-1)^g \sum_{a=1}^N d_a = \frac{1}{2}(-1)^g d_*, \quad g = 1, 2, \quad (7.7.7)$$

where

$$d_* \equiv \frac{1}{N} \sum_a d_a \quad \text{and} \quad y_{..} \equiv \frac{1}{2}(y_{1a} + y_{2a}).$$

Equation (7.7.1) can now be written

$$\tilde{\sigma}_e^2 = \frac{1}{N-1} \sum_{a=1}^N \sum_{g=1}^2 \frac{1}{4} (d_a - d_*)^2 = \frac{1}{2} \frac{N}{N-1} s^2(d_*), \quad (7.7.8)$$

where  $s^2(d_*)$  is the sample variance

$$s^2(d_*) \equiv \frac{1}{N} \sum_{a=1}^N (d_a - d_*)^2. \quad (7.7.9)$$

Finally, by (7.7.5),

$$\tilde{\sigma}_e^2 = \frac{1}{2} \frac{N}{N-1} [s^2(y_{1*}) + s^2(y_{2*}) - 2s(y_{1*}, y_{2*})], \quad (7.7.10)$$

where

$$s(y_{1*}, y_{2*}) \equiv \frac{1}{N} \sum_{a=1}^N (y_{1a} - y_{1*})(y_{2a} - y_{2*}). \quad (7.7.11)$$

When just two forms are available, (7.7.8) or (7.7.10) provides an unbiased estimate, over random sampling of people and tests, for an upper bound to the group specific error variance averaged over all  $G$  tests.

Table 8.4.1 repeats the illustrative data of Table 7.3.1, relabeling  $y_{ga1}$  and  $y_{ga2}$  as  $y_{1a}$  and  $y_{2a}$ . The reader may use (7.7.10) to compute  $\tilde{\sigma}_e^2$  from Table

8.4.1, finding that  $\tilde{\sigma}_e^2 = 9.2$ . By chance, this is less than  $\hat{\sigma}^2(E_{g*}) = 9.9$ , found in Table 7.3.1; usually it would be larger than  $\hat{\sigma}^2(E_{g*})$ .

If, in some case,  $Y_{1*}, Y_{2*}, \dots, Y_{g*}, \dots, Y_{n*}$  were effectively replicate measurements, so that  $\sigma_\alpha^2 = 0$  and  $\sigma(E_{1*}) = \sigma(E_{2*}) = \dots = \sigma(E_{n*})$ , then we could use (7.7.2) and (7.6.8) to show that  $\tilde{\sigma}_e^2$  of (7.7.1) and  $\hat{\sigma}^2(E_{g*})$  of (7.3.16) would be unbiased estimates of the same parameter. Furthermore, since  $\tilde{\sigma}_e^2$  would be based on only  $(N - 1)(n - 1)$  degrees of freedom, the estimate  $\hat{\sigma}^2(E_{g*})$  would be preferred.

### Exercises

- 7.1. Given the data of Table 7.3.1, compute an unbiased estimate of the specific error variance of each examinee.
- 7.2. Given the following set of data, compute the variance components analysis estimates of the error and true-score variances.

Replications $k$	Persons $a$									
	1	2	3	4	5	6	7	8	9	10
1	32	39	48	46	54	57	55	58	69	63
2	33	37	40	49	48	53	50	52	65	61
3	29	35	41	47	51	52	54	59	65	67

Compute these estimates using all three replications.

[Answer:  $\hat{\sigma}^2(E_{g*}) = 8.17$ ,  $s^2(Y_{g**}) = 109.23$ ,  $\hat{\sigma}^2(T_{g*}) = 118.64$ .]

- 7.3. For the data given in the preceding exercise, compute these same estimates using replications 1 and 2 only.  
[Answer:  $\hat{\sigma}^2(E_{g*}) = 10.55$ ,  $s^2(Y_{g**}) = 98.07$ ,  $\hat{\sigma}^2(T_{g*}) = 103.69$ .]
- 7.4. Show that the within-tests sum of squares in Table 7.7.2 is equal to the among-persons sum of squares plus the interaction sum of squares in Table 7.7.1.
- 7.5. Using this result, show that  $\tilde{\sigma}_e^2$  can be expressed as

$$\tilde{\sigma}_e^2 = \frac{N}{(n-1)(N-1)} \left[ \sum_{g=1}^n s^2(y_{g*}) - ns^2(y_{**}) \right],$$

which is formula (7.7.3).

- 7.6. Subtract 12 points from each value of  $y_{ga2}$  in Table 7.3.1. Using these data:
  - a) Obtain two unbiased estimates of  $\sigma^2(T_{g*})$  using formulas (7.5.2) and (7.5.5).
  - b) Compare the estimate obtained using (7.5.2) with the estimate obtained using the original data.
  - c) Show that the estimate obtained using (7.5.5) would be the same if the original data were used.
- 7.7. Refer to the data of Table 7.3.1. Suppose that there are no replications but only two nominally parallel tests (that is,  $y_{1a}$  and  $y_{2a}$  instead of  $y_{ga1}$  and  $y_{ga2}$ ). Find the generous estimate  $\tilde{\sigma}_e^2$ .

- 7.8. Find the expected mean square among persons in Table 7.6.1 by using the following procedure.

a) Show that the sum of squares among persons can be expressed as

$$\frac{1}{r} \sum_{a=1}^N y_{ga+}^2 - \frac{1}{Nr} y_{g++}^2.$$

b) Show that

$$\mathcal{E} \left( \frac{1}{r} \sum_{a=1}^N y_{ga+}^2 \right) = Nr\mu^2 + Nr\sigma^2(T_{g*}) + N\sigma^2(E_{g*})$$

by expressing

$$\sum_{a=1}^N y_{ga+}^2 \quad \text{in terms of } \mu, T_{ga}, \text{ and } E_{gak}$$

and taking the expectation.

c) Express  $y_{g++}$  in terms of  $\mu$ ,  $T_{ga}$ , and  $E_{gak}$  and show that

$$\mathcal{E}(y_{g++}^2/Nr) = Nr\mu^2 + r\sigma^2(T_{g*}) + \sigma^2(E_{g*}).$$

d) Find the required expected mean square by using parts (a), (b), and (c).

### References and Selected Readings\*

- CORNFIELD, J., and J. W. TUKEY, Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, **27**, 907–949.
- DAYHOFF, E., Generalized polykays, an extension of simple polykays and bipolykays. *Annals of Mathematical Statistics*, 1966, **37**, 226–241.
- GRAYBILL, F. A., *An introduction to linear statistical models*, Vol. I. New York: McGraw-Hill, 1961.
- GRAYBILL, F. A., and R. A. HULTQUIST, Theorems concerning Eisenhart's model II. *Annals of Mathematical Statistics*, 1961, **32**, 261–269.
- HILL, B. M., Inference about variance components in the one-way model. *Journal of the American Statistical Association*, 1965, **60**, 806–825.
- HOOKE, R., Some applications of bipolykays to the estimation of variance components and their moments. *Annals of Mathematical Statistics*, 1956, **27**, 80–98.
- LORD, F. M., Statistical inferences about true scores. *Psychometrika*, 1959, **24**, 1–17.
- SCHEFFÉ, H., *The analysis of variance*. New York: Wiley, 1959.
- STANLEY, J. C., Statistical analysis of scores from counterbalanced tests. *Journal of Experimental Education*, 1954–55, **23**, 187–208.

---

\* This list of references has been kept short because the literature deals primarily with reliability and only secondarily or not at all with error variance, and because most of the literature using analysis of variance components has been based on undesirable assumptions that involve the normality and independence of residuals.

- STANLEY, J. C., Analysis-of-variance principles applied to the grading of essay tests. *Journal of Experimental Education*, 1962, **30**, 279-283.
- STANLEY, J. C., Reliability. In R. L. Thorndike (Ed.), *Educational measurement*. Washington: American Council on Education (in preparation).
- THOMPSON, W. A., JR., The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, 1962, **33**, 273-289.
- TIAO, G. C., and W. Y. TAN, Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance-components. *Biometrika*, 1965, **52**, 37-53.
- WINER, B. J., *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Part 3

## OTHER WEAK TRUE-SCORE MODELS



## CHAPTER 8

# SOME TEST THEORY FOR IMPERFECTLY PARALLEL MEASUREMENTS

### 8.1 Defining True Score

Consider an examiner who has obtained one measurement on each of a number of people. If he is perfectly satisfied with his measurements, that is, if he feels that the score of each individual accurately represents the psychological variable that he is trying to measure, then he will see little reason to concern himself with mental test theory and he will proceed to use the scores as they are.

On the other hand, the examiner may feel that these scores may slightly misrepresent the abilities of the individuals being measured. For example, he may feel that the scores might have been different if a different but equally satisfactory test had been used, or if the test had been administered at a different time or under different conditions. In this case, the examiner will be interested in something other than the test scores that he has at hand.

It is helpful if the examiner can clearly define the psychological variable that he intends to study. The examiner should consider all the possible measurements that he might make for his present purpose, and then ask himself whether the psychological variable that he intends to study can be defined in terms of this hypothetical set of measurements. If the variable cannot be defined in these terms, then the information the examiner seeks cannot be found by available measurement techniques. The examiner must then either improve his measurement techniques or decide to study some variable that can be measured by available techniques.

Just how is such a variable to be defined? In Chapters 1 through 7, we assumed that the variable of immediate interest to the examiner can be defined as the expected value of the measurements he has obtained, the expectation being taken over the (hypothetical) set of all parallel measurements. In Chapter 7, we called this expected value the specific true score, to distinguish it from other kinds of true score. The present chapter is concerned not only with situations where it is impossible for practical reasons to obtain parallel measurements, but also with situations where it is undesirable for logical reasons to define true score in terms of any single test form.

The examiner who chooses to study the specific true score is in effect making the following assertion: "If I were allowed more testing time to obtain a *single*

total score for each examinee, I would choose to administer a longer test made up of forms identical (or  $\tau$ -equivalent) to the form actually administered, insofar as this were possible without having experimentally dependent errors of measurement". Most examiners would not really wish to utilize additional testing time in this way, however. They would feel that the "true score" in which they are interested has aspects not covered by the items in the test actually administered, aspects that they would wish to cover if additional testing time were available. If they were able to administer several additional test forms in the additional time, obtaining a single total score, it would not disturb them to know that some of these forms were a little more difficult than others or measured along slightly different dimensions, so long as they were sure that each form measured important aspects or manifestations of the psychological variable under study.

Imperfectly parallel test forms, often loosely called "parallel" in practical work, will here be called *nominally parallel*. This term does not imply any mathematically defined degree of parallelism in any particular respect; it implies merely that from the point of view of the examiner, the true score is to be defined in terms of all nominally parallel forms.

The simplest situation to consider is the one in which the examiner conceives of a pool or population of nominally parallel test forms and defines his true score as the expected score over this population of forms. This true score is the *generic true score*  $\xi$ , already mentioned in (7.6.1). For examinee  $a$ ,

$$\xi_a \equiv \mathcal{E}_g Y_{ga} = \mathcal{E}_k \mathcal{E}_g Y_{gak} = \mathcal{E}_g T_{ga}. \quad (8.1.1)$$

The reader may associate the words "generic" and "specific" with the biological terms "genus" and "species". The generic true score is defined in terms of a whole family of different test forms, not just in terms of a single one. The theory of generic true scores, and the related errors of measurement, is the main concern of Chapters 8 and 11 and of certain sections of Chapter 9.

To illustrate, a professor might define "true achievement in mental test theory course number 530 as taught in 1970" as the examinee's expected score on all the final examinations (constructed according to an appropriate 1970 blueprint) that the professor might administer in 1970, rather than on the particular examination questions actually administered. The correlation between the actual examination score and this "true achievement" score (generic true score) must be of more interest than the correlation between the actual examination score and the corresponding specific true score.

Again, for his own immediate purposes an examiner might define "true spelling ability" as the examinee's expected score on all possible spelling items (of a certain type) based on the words in a certain dictionary. He would not mean to imply that this is the best possible definition of the construct "spelling ability"; he would merely mean that he finds this "true score" convenient to measure and interesting to study, at least until he is able to obtain an unbiased

estimate of some more satisfactorily defined function, which he will then call "spelling ability".

The notion of generic true score is implicit in any approach to the analysis of repeated measurements by analysis of variance components. The model studied in Chapter 7 is a special case in which the test forms effect is assumed to be zero. For many practical applications this simpler model is entirely adequate. The idea of using a generic true score has been developed and is strongly advocated by Cronbach and his associates; the reader might see Rajaratnam (1960), and also Cronbach, Rajaratnam, and Gleser (1963). They do not use the term *generic* but speak of *generalizability*, whereas we use the older term *reliability*.

Sometimes good reasons can be advanced for defining true score as a *weighted* expectation of the observed test scores, taken over all nominally parallel tests. The weighting may well depend on the examiner's subjective judgment. For example, the examiner may well feel that the ability to spell certain unusual or technical words in the dictionary is less important for what he means by "spelling ability" than is the ability to spell other, more usual words. If so, he can think of each single word as a "test" and weight each as seems appropriate for his definition of "spelling ability". Another very reasonable possibility is that the examiner would transform all his observed measurements before taking an expectation. For example, all his observed measurements could be standardized, normalized, or transformed into ranks before averaging.

In each of the foregoing situations,  $\zeta'_a$ , the true score of examinee  $a$ , can be defined as

$$\zeta'_a \equiv \mathcal{E}_g w_g Y_{ga}, \quad (8.1.2)$$

where  $w_g$  is some weight assigned to test  $g$  and  $Y_{ga}$  is any transformed or untransformed measurement. If we define  $Y'_{ga} \equiv w_g Y_{ga}$  and drop all primes, (8.1.2) becomes the same as (8.1.1). Thus all the situations described in the preceding paragraph are covered by the generic true score of (8.1.1), once it is agreed that in (8.1.1)  $Y_{ga}$  represents the (possibly transformed) score whose expected value is the generic true score, rather than the untransformed score that may be the usual "raw score" of the test administered.

The examiner who feels that the generic true score is a quantity of interest needs a treatment of generic true score, and of the corresponding generic error of measurement, similar to that already given for specific true score and specific error of measurement. The purpose of the present chapter is to provide such a treatment, within the limitations of the space available. The reader should refer back to Chapter 7 whenever this will help him to see the similarities and differences between the two developments more clearly. The reader who does not wish to spend further time studying generic true score can skip the remainder of this chapter and the relevant sections of Chapter 9 without loss of continuity.

Of basic importance are the properties of the generic errors of measurement, discussed in Section 8.4. These properties are not as simple as those of the

specific errors, and for this reason they need careful consideration. On the other hand, since the theory does not require that the generic errors on any particular test be unbiased, nor that generic errors be experimentally linearly independent, generic theory covers not only the situations covered by classical test theory, but also many others that have been of concern to practical testers.

Beyond this, the function of Chapter 8, much like that of Chapter 7, is to find ways of estimating the variance of the generic errors of measurement (Sections 8.3 and 8.7) and of the generic true scores (Section 8.5) from actual test data. For the most part, the remaining sections are concerned with comparing different available estimates of generic and specific variances. Some attempt is made to show how the error variances are used in practical work. The real point, however, is that a generic error variance has much the same uses for an examiner who is interested in generic true scores as does a specific error variance for an examiner who is interested in specific true scores.

At first reading, the reader may wish to focus his attention on the principles that require the definition of different error variances. He may leave the detailed methods for estimating these for later study.

## 8.2 The Generic Error of Measurement

If an examiner is interested in the observed score as a measurement that conveys information about the generic true score, then he will frequently be interested in the *generic error of measurement*, defined by

$$\epsilon_{ga} \equiv y_{ga} - \zeta_a, \quad (8.2.1)$$

which is simply *the discrepancy between the measurement at hand ( $y_{ga}$ ) and the measurement of interest ( $\zeta_a$ )*. From earlier chapters, the reader already appreciates some of the practical uses of estimated error variances and other parameters that describe errors. These same practical considerations now lead us to try to find formulas for parameters that describe generic errors.

This chapter deals with the ordinary situation where replicate measurements  $y_{gak}$  and  $y_{gal}$  are not available. The measurements that are actually available instead may be symbolized as  $y_{gak}$  and  $y_{hak}$ , or as  $y_{ga(k)}$  and  $y_{ha(k)}$ , or, dropping the third subscript, simply as  $y_{ga}$  and  $y_{ha}$ .

When  $n$  test forms have been administered to  $N$  examinees, the available data are the scores  $y_{ga}$ ,  $g = 1, 2, \dots, n$ ,  $a = 1, 2, \dots, N$ . According to the general analysis of variance model (7.6.5),

$$Y_{ga(k)} \equiv \mu + (\zeta_a - \mu) + (\pi_g - \mu) + (\alpha_{ga} + E_{ga(k)}).$$

(The two terms  $\alpha_{ga}$  and  $E_{ga(k)}$  cannot be computed separately from each other in the absence of replication.) This model and (8.2.1) lead to the identity

$$\epsilon_{ga} \equiv E_{ga(k)} + \alpha_{ga} + (\pi_g - \mu), \quad (8.2.2)$$

where  $\pi_g \equiv \mathcal{E}_g Y_{ga}$  is the test difficulty,  $\mu \equiv \mathcal{E}_a \zeta_a$  is the mean true score, and  $\alpha_{ga}$  is the test-examinee interaction as defined in Section 7.6. *Equation (8.2.2) relates the generic error of measurement to the specific error of measurement  $E_{ga}$  treated in detail in earlier chapters.* Note that the generic error of measurement includes whatever discrepancies arise from differences between test forms and from interaction between persons and test forms.

Clearly, for examinee  $a$ , the expectation (over tests) of the generic error is zero:

$$\mathcal{E}_g \epsilon_{ga} = \mathcal{E}_g (Y_{ga} - \zeta_a) = \mathcal{E}_g Y_{ga} - \mathcal{E}_g \zeta_a = 0, \quad a = 1, 2, \dots, N. \quad (8.2.3)$$

Thus, if the expectation is taken over tests, the generic error has the same property of unbiasedness that is essential for the development of the basic test theory model of Chapter 2.

Further properties of generic errors will be considered in Section 8.4, but first we shall find an estimate for their variance. Here and throughout this chapter, it is assumed that the population of tests is infinitely large, unless otherwise specified.

### 8.3 The Generic Error Variance

The variance over tests of  $\epsilon_{ga}$  is called the *generic error variance* for examinee  $a$ . The generic error variance tells how much trust can be put in an examinee's observed score as a measure of the generic true score. By (8.2.3), this variance is

$$\sigma^2(\epsilon_{*a}) = \mathcal{E}_g \epsilon_{ga}^2. \quad (8.3.1)$$

Thus the generic error variance differs from the specific error variance only in that an expectation over nominally parallel forms replaces an expectation over replications (strictly parallel forms). The practical formulas for estimating the generic error variance are analogous to those for the specific error variance already presented in Section 7.3, except that summations are taken over tests instead of replications. Since the derivations are similar, the results are stated here in summary form only. We shall give a brief comparison of specific and generic error variances, followed by a practical example.

As in (7.3.9), an estimate of generic error variance from just two *randomly selected* test forms ( $g = 1, 2$ ), unbiased over nominally parallel tests, is

$$\hat{\sigma}^2(\epsilon_{*a}) = \frac{1}{2} d_a^2, \quad (8.3.2)$$

where

$$d_a \equiv y_{2a} - y_{1a} = \epsilon_{2a} - \epsilon_{1a}.$$

We may understand that  $\hat{\sigma}^2(\epsilon_{*a})$  really is unbiased from the following formulation:

$$\begin{aligned} \mathcal{E}_g \mathcal{E}_h \frac{1}{2} (\epsilon_{ga} - \epsilon_{ha})^2 &= \frac{1}{2} \mathcal{E}_g \mathcal{E}_h \epsilon_{ga}^2 + \frac{1}{2} \mathcal{E}_g \mathcal{E}_h \epsilon_{ha}^2 - \mathcal{E}_g \epsilon_{ga} \mathcal{E}_h \epsilon_{ha} \\ &= \mathcal{E}_g \epsilon_{ga}^2 = \sigma^2(\epsilon_{*a}). \end{aligned}$$

In general, as in (7.3.13), given  $n$  randomly selected test forms,

$$\hat{\sigma}^2(\epsilon_{*a}) = \frac{n}{n-1} s^2(y_{*a}) \quad (8.3.3)$$

is an estimate, unbiased with respect to random sampling of tests, for the generic error variance of examinee  $a$ , where

$$s^2(y_{*a}) \equiv \frac{1}{n} \sum_{g=1}^n (y_{ga} - y_{*a})^2. \quad (8.3.4)$$

By a fundamental identity in analysis of variance (see Theorem 2.6.2), the variance  $\sigma^2(\epsilon_{**}) = \mathcal{E}_g \mathcal{E}_a \epsilon_{ga}^2$  may be written

$$\sigma^2(\epsilon_{**}) = \mathcal{E}_a \sigma^2(\epsilon_{*a}) + \sigma_a^2(\mathcal{E}_g \epsilon_{ga}) = \mathcal{E}_a \sigma^2(\epsilon_{*a}). \quad (8.3.5)$$

The last result follows because  $\mathcal{E}_g(\epsilon_{ga}) \equiv 0$  for every  $a$  and thus  $\sigma_a^2(\mathcal{E}_g \epsilon_{ga}) = 0$  also. The group generic error variance is defined as  $\mathcal{E}_a \sigma^2(\epsilon_{*a})$ , that is, as the average over examinees of the generic error variance for examinee  $a$ . Because of (8.3.5) we denote it by the symbol  $\sigma^2(\epsilon_{**})$ .

We may formulate an estimate of the group generic error variance, unbiased over random sampling of people and tests:

$$\hat{\sigma}^2(\epsilon_{**}) = \frac{1}{N} \sum_{a=1}^N \hat{\sigma}^2(\epsilon_{*a}), \quad (8.3.6)$$

that is, the average of the unbiased estimates for the  $N$  examinees in the sample. That  $\hat{\sigma}^2(\epsilon_{**})$  really is unbiased we can see from the following formulation, obtained from (8.3.6), (8.3.3), and (8.3.5):

$$\mathcal{E}_g \mathcal{E}_a \hat{\sigma}^2(\epsilon_{**}) = \frac{1}{N} \sum_{a=1}^N \mathcal{E}_a \mathcal{E}_g \hat{\sigma}^2(\epsilon_{*a}) = \frac{1}{N} \sum_{a=1}^N \mathcal{E}_a \sigma^2(\epsilon_{*a}) = \sigma^2(\epsilon_{**}).$$

A computational formula for the unbiased estimate is

$$\hat{\sigma}^2(\epsilon_{**}) = \frac{1}{N(n-1)} \sum_{a=1}^N \sum_{g=1}^n (y_{ga} - y_{*a})^2. \quad (8.3.7)$$

We have assumed here that the same randomly selected forms have been used for each examinee. Formula (8.3.6) would still provide an unbiased estimator even if a different set of  $g = 1, 2, \dots, n$  forms were randomly selected for each examinee.

When just two tests have been administered,

$$\hat{\sigma}^2(\epsilon_{**}) = \frac{1}{2N} \sum_{a=1}^N d_a^2. \quad (8.3.8)$$

This is the same as (7.3.12), the corresponding estimated specific error variance, except that  $d_{ga}$  has been replaced by  $d_a$ , the difference between observed scores

on two randomly selected forms. We may make similar comparisons between (8.3.7) and (7.3.17), between (8.3.6) and (7.3.16), and between (8.3.5) and (2.6.2). If  $y_{ga1}$  and  $y_{ga2}$  in the numerical example of Table 7.3.1 are replaced by  $y_{1a}$  and  $y_{2a}$  and if summations over the third subscript are replaced by summations over the first subscript, the computations outlined there lead to an estimated generic error variance  $\hat{\sigma}^2(\epsilon_{**}) = 9.9$  instead of an estimated specific error variance  $\hat{\sigma}^2(E_{g*}) = 9.9$ . This numerical identity is the natural result of the fact that not the numbers but only their interpretation has been changed.

We see from (8.2.2) and (7.6.8) that *the relation between the group generic error variance and the group specific error variance is*

$$\begin{aligned}\sigma^2(\epsilon_{**}) &= \sigma^2(E_{**}) + \sigma^2(\alpha_{**}) + \sigma^2(\pi_*) \\ &= \mathcal{E}_g \sigma^2(E_{g*}) + \sigma^2(\alpha_{**}) + \sigma_\pi^2.\end{aligned}\quad (8.3.9)$$

The variances on the right-hand side of this equation are, respectively:

- 1) the variance due to replications, that is, the specific error variance averaged over all nominally parallel test forms (see Eq. 7.6.8),
- 2) the interaction variance

$$\sigma^2(\alpha_{**}) = \mathcal{E}_g \mathcal{E}_a \alpha_{ga}^2, \quad (8.3.10)$$

- 3) the among-tests variance.

The additive form of this equation results from the construction of uncorrelated person, test, interaction, and error variables.

From (8.3.9), we see that *the generic error variance is greater than or equal to the specific error variance averaged over the population of nominally parallel test forms. Any difference in difficulty between the nominally parallel forms, and also any interaction between persons and forms, makes the generic variance larger than the mean specific variance.*

As an example of the use of the generic error variance, consider the Junior Professional Assistant (*JPA*) Examination of some state civil service commission. In periods of rapid hiring, different nominally parallel forms of this examination are administered week after week. It has been found repeatedly that the scores of an examinee who takes several randomly chosen forms of the examination show no practice effect and, further, appear to have an approximately normal distribution, with standard deviation  $\sigma(\epsilon_{*a}) = 3.00$ .

Suppose that Mary Smith takes one (randomly chosen) form of the examination and receives a score of 65.00 (such scores are reported to two decimal places). The minimum passing grade on the examination has been fixed by law at 70.00. Assuming that she does not change in the interim, what light can we throw on Mary's chances of obtaining a score of at least 70.00 if she takes a second (randomly chosen) form of the test?

If Mary's first and second scores are drawn at random from a normal distribution with unspecified mean and standard deviation of 3, their difference is normally distributed with zero mean and a standard deviation of  $3\sqrt{2}$ . Reference to standard normal-curve tables shows that such a difference will exceed  $5/(3\sqrt{2}) = 1.18$  standard deviation units approximately 24% of the time. Thus, for a randomly chosen examinee whose first score is 65.00, there is a 12% chance that the second score will be no more than  $65 - 5 = 60$  and a 12% chance that it will equal or exceed the passing score of  $65 + 5 = 70$ .

#### 8.4 Basic Properties of Generic Errors of Measurement

Here and elsewhere, unless otherwise specified, we assume the usual situation in which each test form is administered to every examinee. The unusual case in which each examinee takes a different form of the test has been covered in the preceding section and will be mentioned again from time to time.

**Bias.** It has already been shown (Eq. 8.2.3) that for a given examinee  $a$ , the expected generic error is zero in the population of tests:

$$\mathcal{E}_g \epsilon_{ga} = 0, \quad a = 1, 2, \dots, N.$$

Note, however, that the test difficulty  $\pi_g \equiv \mathcal{E}_a Y_{ga}$  defined by (7.6.3) does not ordinarily equal the mean generic true score  $\mu \equiv \mu_\xi \equiv \mathcal{E}_a \xi_a = \mathcal{E}_g \mathcal{E}_a Y_{ga} = \mathcal{E}_g \pi_g$  (Eq. 7.6.2). Consequently, for a single fixed test  $g$ , the expectation over examinees of the generic error, that is,

$$\mathcal{E}_a \epsilon_{ga} = \mathcal{E}_a (Y_{ga} - \xi_a) = \pi_g - \mu_\xi, \quad (8.4.1)$$

is not ordinarily zero in the population of examinees. For a single fixed test  $g$ , the generic errors of measurement are usually biased; the errors tend to be negative or positive, depending on whether test  $g$  is more difficult or less difficult than the tests nominally parallel to it.

If some nominally parallel forms of the Junior Professional Assistant Examination are more difficult than others, this clearly is unfair to the examinees taking these forms. A possible correction procedure would be to subtract an estimate of  $\pi_g - \mu$  from the score of each examinee. An unbiased estimate of the desired correction factor  $\pi_g - \mu$  can be obtained by administering the available nominally parallel test forms to large random samples from an appropriate population (see Section 8.7). Such a correction is, precisely, the first step toward equating nominally parallel test forms. It does not solve the entire equating problem, as we shall point out a few paragraphs hence.

**Correlation of errors over testings.** When nominally parallel test forms are of unequal difficulty, the generic errors of measurement for two examinees  $a$  and  $b$  tend to be positively correlated over forms. Examinees tend to get low scores on difficult forms, high scores on easy ones.

**Table 8.4.1**

Modification of Table 7.3.1 to illustrate the computation of estimates of generic error variance, true-score variance, and reliability

$g \backslash a$	1	2	3	4	5	6	7	8	9	10
$y_{ga}$	125	119	109	104	101	98	97	94	90	81
$y_{ga}$	2	120	122	107	108	98	106	96	99	87

$\hat{\sigma}^2(\epsilon_{**}) = 9.900$	$s(y_{1*}, y_{2*}) = 126.92$
$s^2(y_{**}) = 0.81$	$\hat{\sigma}_\zeta^2 = 141.02$
$\tilde{\sigma}_e^2 = 9.2$	$s^2(y_{**}) = 131.06$
$\hat{\sigma}_\pi^2 = 0.70$	

More explicitly, consider the expectation, taken over all possible pairs of examinees  $a, b = 1, 2, \dots, \infty$ , of the covariance between  $\epsilon_{ga}$  and  $\epsilon_{gb}$ , taken over tests:

$$\begin{aligned} \mathcal{E}_a \mathcal{E}_b \sigma(\epsilon_{*a}, \epsilon_{*b}) &= \mathcal{E}_a \mathcal{E}_b \mathcal{E}_g(Y_{ga} - \zeta_a)(Y_{gb} - \zeta_b) \\ &= \mathcal{E}_g \mathcal{E}_a(Y_{ga} - \zeta_a) \mathcal{E}_b(Y_{gb} - \zeta_b) \\ &= \mathcal{E}_g (\pi_g - \mu_\zeta)^2 = \sigma_\pi^2. \end{aligned} \quad (8.4.2)$$

(If two examinees are drawn at random with replacement from an infinite population, the probability that  $a$  and  $b$  are the same person is zero; consequently the second equality holds.) The derivation shows that this expected covariance must always be positive unless all test forms are of equal difficulty.

A formula for an unbiased sample estimate of  $\sigma_\pi^2$  is readily obtainable from the among-tests expected mean square of Table 7.7.1:

$$\hat{\sigma}_\pi^2 = \frac{n}{n-1} s^2(y_{**}) - \frac{1}{N} \tilde{\sigma}_e^2.$$

Table 8.4.1 repeats the illustrative data of Table 7.3.1, relabeling  $y_{ga1}$  and  $y_{ga2}$  as  $y_{1a}$  and  $y_{2a}$ . The reader will find  $\hat{\sigma}_\pi^2$  for Table 8.4.1 to be

$$\hat{\sigma}_\pi^2 = (2)(0.81) - (9.2/10) = 0.70.$$

As an example of one effect of this correlation of errors, consider the practical problem mentioned at the end of Section 7.4. On a certain test administration, John's score is  $y_{ga} = 91$  and Mary's is  $y_{gb} = 98$ . What can we say of the possibility that they are of virtually equal ability?

If the errors were uncorrelated over testings (as they were in Chapter 7), the sampling variance of  $Y_{*a} - Y_{*b}$  over testings would be simply  $\sigma^2(\epsilon_{*a}) + \sigma^2(\epsilon_{*b})$ . Since generic errors ordinarily are correlated over nominally parallel testings, the sampling variance is  $\sigma^2(\epsilon_{*a}) + \sigma^2(\epsilon_{*b}) - 2\sigma(\epsilon_{*a}, \epsilon_{*b})$ , which is a possibly much smaller quantity, depending on the size of  $\sigma_\pi^2$ . Thus our willing-

ness to believe that John and Mary may be of virtually equal ability will in many cases depend on whether the term "ability" is interpreted as referring to the specific true score on the test actually administered, or to the generic true score defined by a larger universe of test items.

**Correlation of errors and true scores.** Consider the following breakdown of the observed-score variance:

$$\sigma^2(Y_{g*}) = \sigma^2(\xi_* + \epsilon_{g*}) = \sigma_\xi^2 + 2\sigma(\xi_*, \epsilon_{g*}) + \sigma^2(\epsilon_{g*}).$$

To keep matters simple, suppose for a moment that each test is scaled so that  $\sigma(Y_{g*})$  is the same for each test  $g = 1, 2, \dots$ , the true score being the expected value over tests of the (scaled)  $Y_{ga}$  values. It is now clear from the breakdown given that if two tests have different generic error variances  $\sigma^2(\epsilon_{g*})$ , they must have different values of  $\sigma(\xi_*, \epsilon_{g*})$ . Thus the generic errors of measurement, on at least one of the two tests, must be correlated with generic true score. The important point here is that in general, regardless of any preliminary superpositions, *the generic error of measurement for a single test  $g$  may be either positively or negatively correlated with generic true score*. In the present notation,  $\rho(\epsilon_{g*}, \xi_*)$ , the correlation over people between generic error and generic true score, cannot be assumed to vanish for a particular test  $g$ .

If some nominally parallel forms of the Junior Professional Assistant Examination are more reliable than others, once again a situation exists that is unfair to some examinees. A good examinee who takes an unreliable test is put at a disadvantage because it is easier for him to obtain a high score on a reliable test than on a nominally parallel but relatively unreliable test. This type of unfairness cannot be satisfactorily corrected by any equating procedure, that is, by any rescaling of the test scores. (In principle, unequally reliable tests can never be satisfactorily equated. This seldom-recognized fact must be true, since otherwise there would be no need for any test to be even moderately reliable.)

*When scores from all tests are pooled, there is no overall correlation between generic true score and generic error of measurement.* We can see that this is true because, by (8.2.3),

$$\sigma(\epsilon_{**}, \xi_*) = \mathcal{E}_a \mathcal{E}_g \epsilon_{ga} (\xi_a - \mu_\xi) = \mathcal{E}_a (\xi_a - \mu_\xi) \mathcal{E}_g \epsilon_{ga} = 0. \quad (8.4.3)$$

*For a single test  $g$ , observed-score variance does not in general equal generic true-score variance plus generic error variance, since*

$$\sigma^2(Y_{g*}) = \sigma^2(\xi_* + \epsilon_{g*}) = \sigma^2(\xi_*) + \sigma^2(\epsilon_{g*}) + 2\sigma(\epsilon_{g*}, \xi_*), \quad (8.4.4)$$

and the last term does not usually vanish. When scores from all tests are pooled, however, the covariance term does vanish, by (8.4.3), so that

$$\sigma^2(Y_{**}) = \sigma^2(\xi_* + \epsilon_{**}) = \sigma^2(\xi_*) + \sigma^2(\epsilon_{**}), \quad (8.4.5)$$

where by definition

$$\sigma^2(Y_{**}) \equiv \mathcal{E}_g \mathcal{E}_a (Y_{ga} - \mathcal{E}_g \mathcal{E}_a Y_{ga})^2. \quad (8.4.6)$$

*When all tests are pooled, observed-score variance does equal generic true-score variance plus generic error variance.*

The fact that the generic error of measurement for a particular test  $g$  may be correlated with generic true score means, first of all, that the properties of test  $g$  as a measuring instrument cannot be summarized in a single reliability coefficient; that is, no single coefficient can summarize the relation between the observed measurement on test  $g$  and the generic true measurement. This is true because the ratio of generic true-score variance to observed-score variance for test  $g$  is no longer equal to the squared correlation of generic true score with observed score on test  $g$ . The changes in our conceptual scheme that are necessary to deal with this situation are briefly discussed in Section 9.8. A formula for estimating the covariance between generic error on a particular test and generic true score is given by (8.7.11).

The fact that different nominally parallel test forms may differ in discriminating power means that it may be insufficient to consider parameters (such as the generic error variance) that describe a whole group of such tests. It may be necessary to describe each nominally parallel test form by parameters computed specially for it. Ways of doing this are treated briefly in Section 8.7.

**Correlation of errors over examinees.** Finally, consider the covariance over examinees between the generic errors of measurement on test  $g$  and those on test  $h$ . This covariance does not usually vanish for a given pair of tests. On the average over an infinite population of tests, however, we see by (8.2.3) that

$$\begin{aligned} \mathcal{E}_g \mathcal{E}_h \sigma(\epsilon_{g*}, \epsilon_{h*}) &= \mathcal{E}_g \mathcal{E}_h \mathcal{E}_a [\epsilon_{ga} - \mathcal{E}_a(\epsilon_{ga})][\epsilon_{ha} - \mathcal{E}_a(\epsilon_{ha})] \\ &= \mathcal{E}_a \mathcal{E}_g [\epsilon_{ga} - \mathcal{E}_a(\epsilon_{ga})] \mathcal{E}_h [\epsilon_{ha} - \mathcal{E}_a(\epsilon_{ha})] \\ &= 0. \end{aligned} \quad (8.4.7)$$

[The second equality holds for a reason similar to that given after (8.4.2)]. *The covariance over examinees of the generic errors of measurement on two tests  $g$  and  $h$  will in general be nonzero. However, the expected value of such covariances over all pairs of tests in an infinite population of tests will be zero.*

Since the correlation between nominally parallel forms will differ from form to form, the correlation between two given forms cannot be used as the reliability coefficient of either. Some of the problems that arise from this fact are treated in Section 9.8.

**Summary.** In this section, we have shown that for a particular test  $g$ ,

- 1) the generic errors may be biased ( $\mathcal{E}_a \epsilon_{ga} \neq 0$ ),
- 2) generic error may correlate with generic true score,

- 3) generic error on one form may correlate (over examinees) with generic error on any other given nominally parallel form, and
- 4) the generic errors of two examinees may be experimentally linearly dependent (see Section 2.11) in repeated testings with different test forms.

These facts considerably complicate the theory of generic true scores. However, these same complications make generic true-score theory applicable in situations that have troubled psychometricians—situations in which the psychometrician insists on considering that his errors of measurement are biased, situations in which he knows that the errors of measurement are not experimentally linearly independent over repeated testings.

### 8.5 Generic True-Score Variance

The generic true score of examinee  $a$  is defined by (8.1.1). The *generic true-score variance* is

$$\sigma_{\xi}^2 \equiv \sigma^2(\xi_*) \equiv \mathcal{E}_a(\xi_a - \mathcal{E}_a \xi_a)^2 = \mathcal{E}_a \xi_a^2 - (\mathcal{E}_a \xi_a)^2. \quad (8.5.1)$$

The symbols  $\sigma_{\xi}^2$  and  $\sigma^2(\xi_*)$  have the same meaning; they are used interchangeably, the former for simplicity, the latter for explicitness.

If  $\sigma_{\xi}^2$  is zero, examinees at hand do not differ on  $\xi$ , so there is probably little interest in trying to measure  $\xi$  for these examinees. If  $\sigma_{\xi}^2$  is a sizable fraction of the observed score variance, however, then clearly there is something to be measured, something that is not being completely drowned in noise (errors of measurement). The main practical use of  $\sigma_{\xi}^2$  is in the ratios  $\sigma_{\xi}^2/\sigma_Y^2$  and  $\sigma_{\xi}^2/\sigma_E^2$ . These are to be discussed in the next chapter, but methods for estimating  $\sigma_{\xi}^2$  from actual test data are most conveniently noted here.

A formula for estimating the generic true-score variance is obtained from the components analysis shown in Table 7.7.1. This is the same as Table 7.6.1 except that there are no replications (that is,  $r = 1$ ). From the last column of Table 7.7.1 we see that when  $G = \infty$ , the expected mean square for people differs from that for interaction by  $n\sigma_{\xi}^2$ . Consequently we may write an estimate for the generic true-score variance that is unbiased over simultaneous random sampling of people and tests:

$$\hat{\sigma}^2(\xi_*) \equiv \hat{\sigma}_{\xi}^2 \equiv \frac{1}{n} \left[ \frac{nN}{N-1} s^2(y_{..}) - \hat{\sigma}_e^2 \right], \quad (8.5.2)$$

where  $\hat{\sigma}_e^2$  is the interaction mean square defined by (7.7.1) and reformulated in (7.7.3). Substituting from (7.7.3) for  $\hat{\sigma}_e^2$  in (8.5.2), we have

$$\hat{\sigma}_{\xi}^2 = \frac{N}{(n-1)(N-1)} \left[ ns^2(y_{..}) - \frac{1}{n} \sum_{g=1}^n s^2(y_{g*}) \right]. \quad (8.5.3)$$

When  $n = 2$ ,

$$4s^2(y_{..}) = 4s^2[\frac{1}{2}(y_{1*} + y_{2*})] = s^2(y_{1*}) + s^2(y_{2*}) + 2s(y_{1*}, y_{2*}). \quad (8.5.4)$$

Substituting this in (8.5.3), we have

$$\hat{\sigma}_{\zeta}^2 = \frac{N}{N-1} s(y_{1*}, y_{2*}). \quad (8.5.5)$$

When  $n \geq 2$ , it may be shown that

$$\hat{\sigma}_{\zeta}^2 = \frac{2}{n(n-1)} \frac{N}{N-1} \sum_{i>j} s(y_{i*}, y_{j*}). \quad (8.5.6)$$

Equations (8.5.5) and (8.5.6) give estimates of the generic true-score variance that are unbiased under simultaneous random sampling of people and tests. The result (8.5.5) should be compared with (7.5.4) and (7.5.5), which show that

$$Ns(y_{g*1}, y_{g*2})/(N-1)$$

is an unbiased estimate (although in that case not an optimal one) for the specific true-score variance. If we apply (8.5.5) to the data in Table 8.4.1, we find that

$$\hat{\sigma}_{\zeta}^2 = \frac{1}{9}(126.92) = 14.102.$$

Again, as at the end of Section 7.5, it should be noted that although  $\sigma_{\zeta}^2 \geq 0$  and is assumed positive,  $\hat{\sigma}_{\zeta}^2$  can be negative for finite  $N$  and  $n$ .

## 8.6 The Relation between Generic and Specific True-Score Variances

Having read about specific and generic true-score variances, the reader will wonder how they differ. Which variance is larger? An answer can be obtained by rewriting Table 7.6.1 for the case where  $n = 1$  and  $G = \infty$ . When this is done, the second and third rows disappear. The result for a given test  $g$  is shown in Table 8.6.1.

**Table 8.6.1**  
Hypothetical components analysis for  
replications of a single test  $g$

Source	Sum of squares	Degrees of freedom	Expected mean square
Among persons	$r \sum_{a=1}^N (y_{ga*} - y_{g..})^2$	$N - 1$	$\sigma^2(E_{**}) + r\sigma_{\alpha}^2 + r\sigma_{\zeta}^2$
Replications	$\sum_a \sum_{k=1}^r (y_{gak} - y_{ga*})^2$	$N(r - 1)$	$\sigma^2(E_{**})$
Total	$\sum_a \sum_k (y_{gak} - y_{g..})^2$	$Nr - 1$	

Table 8.6.1 is identical with Table 7.5.1 except for the last column. The reason for the difference is that Table 7.5.1 shows the expectations over people and replications, whereas Table 8.6.1 shows the expectations over people, replications, and tests. Thus, from Table 7.5.1,

$$\mathcal{E}_{ak} \frac{r}{N-1} \sum_a (y_{ga\cdot} - y_{g..})^2 = \sigma^2(E_{g*}) + r\sigma^2(T_{g*}), \quad (8.6.1)$$

whereas, from Table 8.6.1,

$$\mathcal{E}_{gak} \frac{r}{N-1} \sum_a (y_{ga\cdot} - y_{g..})^2 = \sigma^2(E_{**}) + r\sigma_\alpha^2 + r\sigma_\xi^2. \quad (8.6.2)$$

Taking the expectation of (8.6.1) over tests and combining the result with (8.6.2), we have

$$\mathcal{E}_g \sigma^2(E_{g*}) + r\mathcal{E}_g \sigma^2(T_{g*}) = \sigma^2(E_{**}) + r\sigma_\alpha^2 + r\sigma_\xi^2. \quad (8.6.3)$$

By (7.6.8), the two error-variance terms are equal, leaving the basic relationship

$$\sigma_\alpha^2 + \sigma_\xi^2 = \mathcal{E}_g \sigma^2(T_{g*}). \quad (8.6.4)$$

*The expected specific true-score variance exceeds the generic true-score variance by an amount equal to the interaction variance.*

It should not seem odd that the (generic) true-score variance for nominally parallel tests tends to be less than the (specific) true-score variance for rigorously parallel tests. Nominally parallel tests have more sources of error variance than rigorously parallel tests. For both kinds of tests, observed-score variance equals error variance plus true-score variance. Given a fixed observed-score variance, the larger the error variance, the smaller the true-score variance. This line of thinking should give the reader an appreciation of the conclusion that

$$\sigma_\xi^2 \leq \mathcal{E}_g \sigma^2(T_{g*}). \quad (8.6.5)$$

*The generic true-score variance is smaller than the specific true-score variance averaged over tests, except in the unusual case where all test forms are strictly parallel.*

It is apparent from the derivation in Section 7.7 that when  $\sigma_\alpha^2 = 0$ , the generous estimate  $\hat{\sigma}_e^2$  of the specific error variance is actually an unbiased estimate of  $\mathcal{E}_g \sigma^2(E_{g*})$ . Similarly it is apparent from (8.6.4) that when  $\sigma_\alpha^2 = 0$ ,  $\hat{\sigma}_\xi^2$  is actually an unbiased estimate of  $\mathcal{E}_g \sigma^2(T_{g*})$ . This suggests that we may use  $\hat{\sigma}_\xi^2$  when we wish to obtain an estimate of the specific true-score variance in the very common situations where repeated measurements are not rigorously parallel, but are still not too dissimilar.

To summarize:  $\hat{\sigma}_\xi^2$  of (8.5.3) and (8.5.6) may be used as an estimate, unbiased over random samples of people, for a lower bound to the expected specific true-score variance  $\mathcal{E}_g \sigma^2(T_{g*})$ .

### 8.7 Estimating Generic Parameters Describing a Single Test Form

The group generic error variance  $\sigma^2(\epsilon_{..})$  describes the entire genus (population) of nominally parallel tests. It is therefore very useful as a descriptive parameter in situations where different examinees have taken different randomly assigned test forms, or where different subjects have been rated by different randomly assigned judges. In such situations, the data are called *unmatched*. When all examinees have taken a single test form  $g$ , or been rated by a single judge, the data are called *matched*. This terminology has been provided by Rajaratnam (1960), who called attention to the importance and implications of the distinction.

The variance of the generic errors on a particular test  $g$  is denoted by

$$\sigma^2(\epsilon_{ga}) \equiv \mathcal{E}_a \epsilon_{ga}^2 - (\mathcal{E}_a \epsilon_{ga})^2 = \mathcal{E}_a \epsilon_{ga}^2 - (\pi_g - \mu)^2. \quad (8.7.1)$$

By (8.3.5), (8.3.1), and (8.7.1),

$$\sigma^2(\epsilon_{..}) = \mathcal{E}_a \mathcal{E}_g \epsilon_{ga}^2 = \mathcal{E}_g \sigma^2(\epsilon_{ga}) + \sigma_\pi^2.$$

Clearly  $\sigma^2(\epsilon_{..})$  may be large simply because the nominally parallel tests differ in difficulty, and this can happen even though each test has a small generic error variance. Conversely  $\sigma^2(\epsilon_{ga})$ , the variance of the generic errors on test  $g$ , may be small even though  $\sigma^2(\epsilon_{..})$  is large. Since the examiner is ordinarily concerned with the measurement properties of a single test actually administered, there is need for parameters that specifically describe this test.

In Section 8.3, we made use of the generic standard error of measurement  $\sigma(\epsilon_{..}) = 3.00$  to throw some light on Mary Smith's chances of passing the JPA Examination on a second attempt. Suppose, however, that we obtain some additional information; suppose that the generic standard error of measurement for the particular test form that Mary took was not 3.00 but (to take an extreme case)  $\sigma(\epsilon_{ga}) = 10.00$  for each examinee. Clearly our interpretation of Mary's score of 65 is somewhat changed, as, indeed, is our attitude toward the entire group of people hired because of their scores on this test form. We have less confidence in our appraisal of Mary's ability; we suspect that hiring on the basis of this particular test form may not have been very much better than a random selection.

The main purpose of the present section is to find a way to estimate the generic error variance of a particular test form. However, as a first step that will clarify the type of statistical inference involved, consider the simpler but equally important problem of estimating  $\epsilon_{ga}$ , the generic error of examinee  $a$  on test  $g$ .

**Estimating the bias of test  $g$ .** Consider the difference  $d_a$  between the scores of examinee  $a$  on tests  $g$  and  $h$ :

$$d_a \equiv y_{ga} - y_{ha} = (\zeta_a + \epsilon_{ga}) - (\zeta_a + \epsilon_{ha}) = \epsilon_{ga} - \epsilon_{ha}. \quad (8.7.2)$$

Taken as a random variable over the population of tests, this difference is denoted by  $D_a$ . Now is  $d_a$  an unbiased estimate of  $\epsilon_{ga}$ ? If the expectation  $\mathcal{E}_G$  is taken over all possible pairs of tests  $g$  and  $h$ , it is clear that

$$\mathcal{E}_G D_a \equiv \mathcal{E}_g \epsilon_{ga} - \mathcal{E}_h \epsilon_{ha} = 0. \quad (8.7.3)$$

This result seems to prove that  $d_a$  is not the desired unbiased estimate of  $\epsilon_{ga}$ .

However, since estimates are to be obtained for a particular test  $g$ , let us hold this test  $g$  fixed and suppose that expectations are taken over all ( $G = \infty$ ) other nominally parallel tests. This type of expectation is denoted by  $\mathcal{E}_{h \cdot g}$ . Since we are considering an infinite population of tests,  $\mathcal{E}_{h \cdot g} \epsilon_{ha} = 0$ . Since  $\mathcal{E}_{h \cdot g} \epsilon_{ga} = \epsilon_{ga}$ , we now see from (8.7.2) that

$$\mathcal{E}_{h \cdot g} D_a = \epsilon_{ga}. \quad (8.7.4)$$

Thus, for a randomly selected test  $h$ ,  $d_a$  is an estimate of  $\epsilon_{ga}$  that is unbiased over nominally parallel tests when test  $g$  is held fixed.

Clearly the properties of an estimate depend on the sampling rule. When estimating the properties of test  $g$  as a measuring instrument, we must consider test  $g$  fixed while we sample randomly over all other tests in the infinite population of nominally parallel tests. *This type of rule for sampling tests is assumed throughout the remainder of this section unless otherwise stated.*

As we noted at the start of Section 8.4, an important quantity for describing test  $g$  is the bias in its generic errors of measurement. By definition (8.4.1), this bias is

$$\mathcal{E}_a \epsilon_{ga} = \pi_g - \mu,$$

the difference between the difficulty of test  $g$  and the average difficulty of all nominally parallel tests. Intuitively one might feel that  $y_{g \cdot} - y_{\cdot \cdot}$  is an unbiased estimate of  $\pi_g - \mu$ ; however, this is incorrect for the type of expectation used here since

$$\mathcal{E}_{h \cdot g} \mathcal{E}_a (y_{g \cdot} - y_{\cdot \cdot}) = \frac{n-1}{n} (\pi_g - \mu).$$

The desired estimate of  $\pi_g - \mu$  is  $y_{g \cdot} - y'_{\cdot \cdot}$ , where

$$y'_{\cdot \cdot} \equiv \frac{1}{n'N} \sum_{h \neq g} \sum_{a=1}^N y_{ha}, \quad (8.7.5)$$

the prime on the  $y$  indicating that the average is taken over  $n' \equiv n - 1$  randomly selected, nominally parallel tests after excluding test  $g$ . Thus, *where test  $g$  is held fixed, the sample quantity  $y_{g \cdot} - y'_{\cdot \cdot}$  provides an estimate of the “bias” in the errors of measurement of test  $g$  (that is, an estimate of  $\mathcal{E}_a \epsilon_{ga}$ ) that is unbiased over random sampling of people and of nominally parallel tests.* The proof is that

$$\mathcal{E}_a \mathcal{E}_{h \cdot g} (y_{g \cdot} - y'_{\cdot \cdot}) = \pi_g - \mu = \mathcal{E}_a \epsilon_{ga}. \quad (8.7.6)$$

If we apply this formula to the (rather limited) data in Table 8.4.1, we find that the estimated bias of test 1 is  $-1.8$ ; that of test 2,  $+1.8$ .

**Estimating the generic error variance for test  $g$ .** We come now to the more difficult problem with which this section began, namely, the problem of estimating  $\sigma^2(\epsilon_{g*})$ , the group generic error variance for test  $g$ . To do this, we must administer test  $g$  to a random sample of  $N$  examinees and must also obtain on the same examinees  $n' \equiv n - 1$  measurements that are nominally parallel to each other and to test  $g$ . We may then obtain the desired estimates by setting up an appropriate analysis of variance components similar to that in Table 7.7.1 but differing from it in two important ways.

1. In obtaining the expected mean squares, we hold test  $g$  fixed while taking expectations over people and over the other nominally parallel tests.
2. We split the “tests” effect and the interaction each into two orthogonal, that is, uncorrelated, parts.

Table 8.7.1 is written for the case where the number of tests in the population of nominally parallel tests is infinite. The symbols in this table are all familiar except that

$$y'_{*a} \equiv \frac{1}{n'} \sum_{h \neq g} y_{ha}. \quad (8.7.7)$$

Only the last two rows in the body of the table are used here. Examination of the last column shows that *where test  $g$  is held fixed, the quantity*

$$\hat{\sigma}^2(\epsilon_{g*}) \equiv \frac{n}{n'} \overline{MSI}_g - \frac{1}{n'} \hat{\sigma}_e^2 \quad (8.7.8)$$

*is an estimate, unbiased over random sampling of examinees and of nominally parallel tests, for the variance of the generic errors in the scores on test  $g$ , where  $\overline{MSI}_g$  is the interaction mean square for test  $g$  and  $\hat{\sigma}_e^2$  is the interaction mean square for the remaining tests.* Since the mean square  $\hat{\sigma}_e^2$  has no degrees of freedom when  $n' = 1$ , it follows that we must administer at least  $n' = 2$  tests in addition to test  $g$  to obtain an unbiased estimate of the group generic error variance of test  $g$  from Table 8.7.1.

The reader should not forget that the bias  $\pi_g - \mu$  in test  $g$  is not reflected in the value of  $\sigma^2(\epsilon_{g*})$ . The bias may be large even though  $\sigma^2(\epsilon_{g*})$  is small, since adding a constant bias to all the errors does not change their variance. Thus it is important to obtain estimates of both  $\pi_g - \mu$  and  $\sigma^2(\epsilon_{g*})$  and to consider them jointly when evaluating the measurements produced by test  $g$ .

**Estimating the covariance between error and true score for test  $g$ .** In generic true-score theory, it is not enough to estimate the generic error variance. One also needs to know to what extent the generic errors of measurement on test  $g$  are correlated with generic true score. This problem was raised and briefly discussed in Sec-

Table 8.7.1

Components analysis showing the orthogonal components due to test  $g$ 

Source	Sum of squares	Degrees of freedom	Expected mean square
Among persons	$n \sum_{a=1}^N (y_{\cdot a} - y_{..})^2$	$N - 1$	$\frac{n'}{n} [\sigma^2(E_{**}) + \sigma_\alpha^2 - \sigma^2(\epsilon_{g*})] + \sigma^2(Y_{g*}) + n' \sigma_\xi^2$
Among tests	$\frac{n'}{n} N(y_{\bullet\bullet} - y'_{..})^2$	1	$\frac{1}{n} [\sigma^2(E_{**}) + \sigma_\alpha^2 + n' \sigma^2(\epsilon_{g*}) + N \sigma_\pi^2 + n' N (\pi_g - \mu)^2]$
Other tests	$N \sum_{h \neq g}^N (y_{h\bullet} - y'_{..})^2$	$n' - 1$	$\sigma^2(E_{**}) + \sigma_\alpha^2 + N \sigma_\pi^2$
Interaction	$\frac{n'}{n} \sum_{a=1}^N (y_{ga} - y_{\bullet a} - y'_{\cdot a} + y'_{..})^2$	$N - 1$	$\frac{1}{n} [\sigma^2(E_{**}) + \sigma_\alpha^2 + n' \sigma^2(\epsilon_{g*})]$
Other tests	$\sum_{h \neq g}^N \sum_{a=1}^N (y_{ha} - y_{h\bullet} - y'_{\cdot a} + y'_{..})^2$	$(n' - 1)(N - 1)$	$\sigma^2(E_{**}) + \sigma_\alpha^2$
Total	$\sum_{h=1}^n \sum_{a=1}^N (y_{ga} - y_{..})^2$	$nN - 1$	

tion 8.4. Other things being equal, if the correlation of generic error of measurement with generic true score is positive for test  $g$ , it will be easier for a good examinee to obtain a good score on test  $g$  than if the correlation is negative. This fact demonstrates the importance of this correlation for interpreting examinees' scores on test  $g$ .

Given a fixed observed-score variance  $\sigma^2(Y_{g*})$  and a fixed generic error variance  $\sigma^2(\epsilon_{g*})$ , it is ordinarily better to have a high positive correlation between generic errors and true score than to have a high negative correlation. The reason is that the covariance, and hence the correlation, between test  $g$  and generic true score,

$$\sigma(Y_{g*}, \xi_*) = \sigma(\xi_* + \epsilon_{g*}, \xi_*) = \sigma_\xi^2 + \sigma(\epsilon_{g*}, \xi_*), \quad (8.7.9)$$

will be higher when the last term is positive than when it is negative.

The covariance of interest,  $\sigma(\epsilon_{g*}, \xi_*)$ , can be estimated by making use of (8.4.4), which may be rewritten

$$\sigma(\epsilon_{g*}, \xi_*) = \frac{1}{2}[\sigma^2(Y_{g*}) - \sigma^2(\xi_*) - \sigma^2(\epsilon_{g*})]. \quad (8.7.10)$$

Therefore, when test  $g$  is held fixed,

$$\hat{\sigma}(\epsilon_{g*}, \xi_*) \equiv \frac{1}{2} \left[ \frac{N}{N-1} s^2(y_{g*}) - \hat{\sigma}^2(\xi_*) - \hat{\sigma}^2(\epsilon_{g*}) \right] \quad (8.7.11)$$

is an estimate, unbiased in random sampling of people and of nominally parallel tests, for the covariance between the generic errors on test  $g$  and generic true score. In Eq. (8.7.11),  $\hat{\sigma}^2(\epsilon_{g*})$  is obtained from (8.7.8), and

$$\hat{\sigma}^2(\xi_*) = \frac{N}{(n'-1)(N-1)} \left[ n's^2(y'_{..}) - \frac{1}{n'} \sum_{h \neq g} s^2(y_{h*}) \right]. \quad (8.7.12)$$

This last equation is the same as (8.5.3), derived in the section on generic true-score variance, except that here it is written for just  $n'$  tests, excluding test  $g$ .

In Chapter 9, we shall need an unbiased estimate of  $\sigma(Y_{g*}, \xi_*)$ . We can readily obtain this by placing a caret ( $\hat{\phantom{x}}$ ) on each  $\sigma$  in (8.7.9),

$$\hat{\sigma}(Y_{g*}, \xi_*) = \hat{\sigma}_\xi^2 + \hat{\sigma}(\epsilon_{g*}, \xi_*), \quad (8.7.13)$$

and then substituting from (8.7.11) and (8.7.12). It can be shown by considerable algebra that (8.7.13) is equivalent to

$$\hat{\sigma}(Y_{g*}, \xi_*) = \frac{1}{n'} \frac{N}{N-1} \sum_{h \neq g} s(y_{g*}, y_{h*}). \quad (8.7.14)$$

## 8.8 Comparisons of Estimates of Error Variance

**Conventional formula.** The conventional formula (3.3.8) expresses the specific error variance in the population of examinees in terms of the correlation be-

tween two strictly parallel tests:

$$\sigma_E^2 = \sigma_Y^2[1 - \rho(Y_1, Y_2)]. \quad (8.8.1)$$

In practical work, sample statistics have usually been substituted for population parameters to produce the formula

$$\text{estimated } \sigma_E^2 \equiv s_y^2(1 - r_{12}), \quad (8.8.2)$$

where  $s_y^2$  is obtained by somehow pooling the two quantities  $s^2(y_{1*})$  and  $s^2(y_{2*})$ , and where  $r_{12}$  is the sample correlation between the two parallel forms. [Note that even if test 1 and test 2 were strictly parallel, so that  $\sigma^2(Y_{1*}) = \sigma^2(Y_{2*})$ , the corresponding sample variances  $s^2(y_{1*})$  and  $s^2(y_{2*})$  would ordinarily still be unequal because of sampling fluctuations in the errors of measurement.] None of the usual ways of defining  $s_y^2$  and  $r_{12}$  will make (8.8.2) an unbiased estimate of (8.8.1).

The classical formula provides a consistent,\* though biased, estimate. Unbiasedness is not necessarily a decisive factor, however; unbiasedness is defined in terms of mean error, and a mean is not the only acceptable measure of central tendency.

The sampling variances of classical estimates obtained from different definitions of  $s_y^2$  doubtless differ from each other and from the estimates recommended here only by terms of order  $1/N^2$ . Usually, and particularly in small samples, one cannot say that a certain estimate of a parameter is clearly better than all others. The estimates recommended in Chapters 7 and 8 have been obtained by a commonly accepted rationale that is known to produce minimum-variance unbiased quadratic estimates under certain conditions, as we pointed out at the end of Section 7.6.

**Generous estimate of the specific error variance.** Table 8.8.1 lists several estimates for certain error variances. Let us first compare the conventional formula (8.8.2) with the generous estimate of (7.7.10), which is

$$\tilde{\sigma}_e^2 = \frac{1}{2} \frac{N}{N-1} [s^2(y_{1*}) + s^2(y_{2*}) - 2s(y_{1*}, y_{2*})]. \quad (8.8.3)$$

In the very special case where  $s(y_{1*}) = s(y_{2*}) = s_y$ , (8.8.3) becomes

$$\tilde{\sigma}_e^2 = \frac{N}{N-1} s_y^2(1 - r_{12}). \quad (8.8.4)$$

The right-hand side of this equation is the same, except for the unbiasing factor  $N/(N-1)$ , as that of the conventional formula (8.8.2). In the general case, however, there is no way of defining  $s_y^2$  as a function of  $s(y_{1*})$  and  $s(y_{2*})$  so as to make formula (8.8.2) or (8.8.4) equivalent to (8.8.3).

---

\* For estimates with finite sampling variance, a *consistent estimate* may be defined as one that is asymptotically unbiased as the number of observations becomes large.

**Table 8.8.1**  
Estimates of error variances

Error variance estimated	Replicate measurements required?	Bias	Symbol	Equation	Formula
Specific, for examinee $a$	yes	none	$\hat{\sigma}^2(E_{ga})$	(7.3.13)	$\frac{r}{r-1} s^2(y_{ga*})$
Specific, for group	yes	none	$\hat{\sigma}^2(E_{g*})$	(7.3.16)	$\frac{1}{N} \sum_a \hat{\sigma}^2(E_{ga})$
Same, averaged over tests	yes	none	$\hat{\sigma}^2(E_{**})$	(7.6.7)	$\frac{1}{n} \sum_g \hat{\sigma}^2(E_{g*})$
Same, averaged over tests	no	positive	$\tilde{\sigma}_e^2$	(7.7.1), (7.7.3)	$\overline{MSI}$
Generic, for examinee $a$	no	none	$\hat{\sigma}^2(\epsilon_{*a})$	(8.3.3)	$\frac{n}{n-1} s^2(y_{*a})$
Generic, for group	no	none	$\hat{\sigma}^2(\epsilon_{**})$	(8.3.6)	$\frac{1}{N} \sum_a \hat{\sigma}^2(\epsilon_{*a})$
Same, for fixed test $g$	no	none	$\hat{\sigma}^2(\epsilon_{g*})$	(8.7.8)	$\frac{n}{n'} \overline{MSI}_g - \frac{\tilde{\sigma}_e^2}{n'}$

Finally it should be remembered that  $\tilde{\sigma}_e^2$  in (8.8.3) is an unbiased estimate of an upper bound, not an unbiased estimate of the error variance  $\sigma_e^2$  itself. The simplicity of the conventional formula (8.8.2) is achieved by (1) slighting the problem of inferring population parameters from observed samples, (2) slighting the difficulties that arise when the available tests do not satisfy the equivalence assumptions made in the derivation, and (3) leaving out of consideration all but one of the many possibly appropriate but essentially different definitions of true score and error. The conventional formula need not be discarded as "wrong", but careful thinking may lead us to prefer one of the alternatives listed in Table 8.8.1.

**Group specific error variance.** When two replicate measurements are available, the estimate (7.3.12) for the group specific error variance may be rewritten

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{2}[d_{g*}^2 + s^2(d_{g*})], \quad (8.8.5)$$

where  $d_{g*}$  and  $s^2(d_{g*})$  are the sample mean and variance

$$d_{g*} \equiv \frac{1}{N} \sum_a d_{ga}, \quad s^2(d_{g*}) \equiv \frac{1}{N} \sum_a (d_{ga} - d_{g*})^2. \quad (8.8.6)$$

Since

$$s^2(d_{g*}) = s^2(y_{g*1}) + s^2(y_{g*2}) - 2s(y_{g*1}, y_{g*2}), \quad (8.8.7)$$

Eq. (8.8.5) can be written

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{2}[s^2(y_{g*1}) + s^2(y_{g*2}) - 2s(y_{g*1}, y_{g*2}) + (y_{g*2} - y_{g*1})^2]. \quad (8.8.8)$$

A comparison of the group specific error variance of (8.8.8) with the generous estimate of (8.8.3) is instructive. In the special case where  $Y_1$  and  $Y_2$  in (8.8.3) are parallel measurements, we see that

$$\hat{\sigma}^2(E_{g*}) = \frac{N-1}{N} \tilde{\sigma}_e^2 + \frac{1}{2}(y_{g*2} - y_{g*1})^2.$$

Clearly the estimate  $\hat{\sigma}^2(E_{g*})$  for the group specific error variance differs from the conventional formula (8.8.2) even more than did the generous estimate  $\tilde{\sigma}_e^2$ . Equation (8.8.8) shows that  $\hat{\sigma}^2(E_{g*})$ , the components analysis estimate of the group specific error variance, uses one source of information (with one degree of freedom) that is ignored by all conventional procedures based on (8.8.2): the quantity  $d_g = y_{g*2} - y_{g*1}$ . This quantity, the difference in sample average score between two replications, contributes valid information about the specific error variance not available in  $s^2(y_{g*1})$  or  $s^2(y_{g*2})$  or  $s^2(d_g)$ , provided the test forms are really parallel.

If in some special case the statistician believes that two test forms differ in mean difficulty but are otherwise parallel, then he cannot use  $d_g$  in this way. In this case, he may prefer some form of the conventional formula (8.8.2) to  $\hat{\sigma}^2(E_{g*})$ .

**Group generic error variance.** When  $n = 2$ , the estimate for the group generic error variance may be rewritten from (8.3.8) as

$$\hat{\sigma}^2(\epsilon_{**}) = \frac{1}{2}[s^2(d_*) + d_*^2], \quad (8.8.9)$$

or

$$\hat{\sigma}^2(\epsilon_{**}) = \frac{1}{2}[s^2(y_{1*}) + s^2(y_{2*}) - 2s(y_{1*}, y_{2*}) + (y_{2*} - y_{1*})^2]. \quad (8.8.10)$$

This last equation is the same as the specific estimate in (8.8.8) except that here we have two different test forms, whereas there we had two replicate measurements on the same test form. If the two test forms of (8.8.10) were as alike as replicate measurements, then  $\tilde{\sigma}_e^2$ ,  $\hat{\sigma}^2(\epsilon_{**})$ ,  $\hat{\sigma}^2(E_{1*})$ , and  $\hat{\sigma}^2(E_{2*})$  would all be unbiased estimates of the same population parameter. Since  $\tilde{\sigma}_e^2$  and  $\hat{\sigma}^2(\epsilon_{**})$  have the same expected value when the test forms are parallel, and since the term  $y_{2*} - y_{1*}$  appears in  $\hat{\sigma}^2(\epsilon_{**})$  but not in  $\tilde{\sigma}_e^2$ , we see that differences in difficulty between the two test forms tend to increase the generic estimate  $\hat{\sigma}^2(\epsilon_{**})$  in comparison with the generous estimate  $\tilde{\sigma}_e^2$ .

## 8.9 Substantive Considerations Regarding Choice among Estimates

The error variances discussed in Chapters 7 and 8, together with their formulas, are summarized in Table 8.8.1. Before we can estimate an error variance in any practical application, we must decide what the word "error" is to mean in that particular context. To do this, we must conceive of some hypothetical

population of sets of repeated measurements from which the obtained set of measurements might have been randomly drawn. If we think of these hypothetical sets of repeated measurements as being strictly parallel, then we are concerned with  $\hat{\sigma}^2(E_{ga})$ ,  $\hat{\sigma}^2(E_{g*})$ , or  $\tilde{\sigma}_e^2$ . If we think of these hypothetical sets of repeated measurements as being imperfectly parallel, then we are concerned with  $\hat{\sigma}^2(\epsilon_{*a})$ ,  $\hat{\sigma}^2(\epsilon_{**})$ , or  $\hat{\sigma}^2(\epsilon_{g*})$ .

Since strictly parallel measures are not ordinarily available, the unbiased estimate (7.3.17) for the group specific error variance,

$$\hat{\sigma}^2(E_{g*}) = \frac{1}{N} \frac{1}{r-1} \sum_{a=1}^N \sum_{k=1}^r (y_{gak} - y_{ga*})^2,$$

cannot ordinarily be computed. However, the generous estimate  $\tilde{\sigma}_e^2$  can be used to get an estimated *upper bound* for the group specific error variance averaged over tests.

If we are concerned with nominally parallel test forms, we can use the generic error variance  $\sigma^2(\epsilon_{**})$  to characterize unmatched measurements, that is, the measurements obtained when different examinees take different randomly assigned test forms or when different subjects are rated by different randomly assigned judges. But  $\sigma^2(\epsilon_{g*})$ , the group generic error variance for test (or judge)  $g$ , is necessary to describe the measurements actually obtained on some particular test  $g$ . These two variances may be quite different.

Note that although  $\sigma^2(\epsilon_{**})$  is to be used for interpreting unmatched data, the formulas given here for estimating  $\sigma^2(\epsilon_{**})$  require matched data. Estimates of  $\sigma^2(\epsilon_{**})$  can also be obtained from unmatched data although less efficiently, from a purely statistical point of view. Unmatched data frequently occur in work with ratings, but they are uncommon in ordinary mental testing. For the sake of simplicity, methods for estimating  $\sigma^2(\epsilon_{**})$  from unmatched data are not discussed here. The reader can obtain the appropriate formula by the components analysis approach used throughout this chapter. Or, see Webster (1960) or Cronbach *et al.* (1963).

When  $\hat{\sigma}^2(\epsilon_{g*})$  is reported, the quantity  $y_{g*} - y_{..}$  should also be reported since it provides an unbiased estimate of the bias of the generic errors of measurement for test  $g$ . In addition, the covariance between generic error of measurement on test  $g$  and generic true score should be estimated (Eq. 8.7.11).

## Exercises

- 8.1. Show that  $\mathcal{E}_g[\sigma(\epsilon_{g*}, \zeta_*)] = \sigma(\epsilon_{**}, \zeta_*) = 0$ .
- 8.2. Show that  $\sigma^2(Y_{**}) = \sigma^2(\zeta_*) + \sigma^2(\epsilon_{**})$  by the following procedure:
  - a) Find  $\mathcal{E}_g\sigma^2(Y_{g*})$  by using (8.4.4), (8.4.3), and the result of Exercise 8.1.
  - b) Use (2.6.3) to give the usual analysis-of-variance breakdown of  $\sigma^2(Y_{**})$  and  $\sigma^2(\epsilon_{**})$ .
  - c) Use the results of (a) and (b) to find  $\sigma^2(Y_{**})$ .

- 8.3. Show that (8.5.3) can be expressed as (8.5.6). [Hint: Expand  $s^2(y_{..})$  as in (8.5.4).]
- 8.4. a) Prove that  $y_{g..} - y_{..}$  is not a consistent estimate of  $\pi_g - \mu$  when expectations are taken with test  $g$  held fixed, that is, when  $N$  becomes large and  $n$  is fixed.  
 b) Prove that  $y_{g..} - y'_{..}$  is an unbiased estimate of  $\pi_g - \mu$  when expectations are taken with test  $g$  held fixed.
- 8.5. a) Show that when two replicate measurements are available,  $\hat{\sigma}^2(E_{g..})$  can be expressed as
- $$\hat{\sigma}^2(E_{g..}) = \frac{1}{2}[s^2(d_{g..}) + d_{g..}^2].$$
- [Hint: Show that  $[s^2(d_{g..}) + d_{g..}^2] = \frac{1}{N} \sum_{a=1}^N d_{ga}^2$ .]
- b) Show that when  $n = 2$ ,  $\hat{\sigma}^2(\epsilon_{..})$  can be expressed as  $\hat{\sigma}^2(\epsilon_{..}) = \frac{1}{2}[s^2(d_{..}) + d_{..}^2]$ . Use the same procedure as in (a).
- 8.6. Find  $\hat{\sigma}(\zeta_*)$  for the data in Table 8.4.1, using formula (8.5.2).
- 8.7. Refer to Exercise 7.2. Suppose that there are no replications but three nominally parallel tests. Using the data given there, find  
 a) the variance components analysis estimates of the generic error variance for each examinee, for the group, and for each test,  
 b) the estimated bias in each test,  
 c)  $\hat{\sigma}(\epsilon_{g..}, \zeta_*)$  for each test, and  
 d)  $\hat{\sigma}(Y_{g..}, \zeta_*)$  for each test.
- 8.8. a) Find  $\hat{\sigma}^2(\epsilon_{..})$  for the data in Table 8.4.1, using formula (8.8.10). Compare this estimate with the generous estimate for these data obtained in Exercise 7.4. [See comment following (8.8.10).]  
 b) Repeat part (a) for the data in Table 8.4.1 with ten points subtracted from each  $y_{2a}$ .

### References and Selected Readings\*

- BUROS, O. K., *Schematization of old and new concepts of test reliability based upon parametric models*. New Brunswick, N.J.: Gryphon Press, 1963. (Dittoed).
- BURT, C., Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology*, 1955, **8**, 103–118.
- CRONBACH, L. J., NAGESWARI RAJARATNAM, and GOLDINE C. GLESER, Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, **16**, 137–163.
- GUTTMAN, L., A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255–282.

---

\* Most of these are concerned primarily with reliability and only secondarily with error variances.

- HAGGARD, E. A., *Intraclass correlation and the analysis of variance*. New York: Dryden, 1958.
- HOYT, C., Test reliability estimated by analysis of variance. *Psychometrika*, 1941, **6**, 153-160.
- JACKSON, R. W. B., Reliability of mental tests. *British Journal of Psychology*, 1939, **29**, 267-287.
- JACKSON, R. W. B., Application of the analysis of variance and covariance method to educational problems. Toronto: Department of Educational Research, University of Toronto. Bulletin No. 11, 1941.
- JACKSON, R. W. B., and G. A. FERGUSON, Studies on the reliability of tests. Toronto: Department of Educational Research, University of Toronto. Bulletin No. 12, 1941.
- LINDQUIST, E. F., *Design and analysis of experiments in psychology and education*. Boston: Houghton-Mifflin, 1953.
- LORD, F. M., Nominally and rigorously parallel test forms. *Psychometrika*, 1964, **29**, 335-346.
- MEDLEY, D. M., and H. E. MITZEL, Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand-McNally, 1963, pp. 247-328.
- RAJARATNAM, NAGESWARI, Reliability formulas for independent decision data when reliability data are matched. *Psychometrika*, 1960, **25**, 261-271.
- RAJARATNAM, NAGESWARI, L. J. CRONBACH, and GOLDINE C. GLESER, Generalizability of stratified-parallel tests. *Psychometrika*, 1965, **30**, 39-56.
- TRYON, R. C., Reliability and behavior domain validity: reformulation and historical critique. *Psychological Bulletin*, 1957, **54**, 229-249.
- WEBSTER, H., A generalization of Kuder-Richardson reliability formula 21. *Educational and Psychological Measurement*, 1960, **20**, 131-138.

# TYPES OF RELIABILITY COEFFICIENTS AND THEIR ESTIMATION

## 9.1 Introduction

When we have a sizable sample of test scores to interpret or to analyze, the group standard error of measurement (the square root of group error variance) is usually a meaningful and useful quantity. It provides information about the degree of inaccuracy of the scores at hand (see Section 7.4). If no actual test scores and no other information about the test is at hand, however, it is a quite uninformative quantity. To say that a score on test  $Y$  has a standard error of measurement of 3 does not by itself lead to any useful conclusion whatever about the general value of the score. For example, one cannot tell whether 3 is unduly large or desirably small without knowing whether examinees commonly differ among themselves by more than 3 score points or by less. In such situations, one requires a reliability coefficient or some similar index, as suggested below.

The standard error of measurement is a dimensional quantity that is expressed in the same units as the observed score. If the observed score is multiplied by a constant, the standard error of measurement is automatically multiplied by the same constant. When we are describing a test, and especially when we are describing the general value of the score as mentioned above, it is helpful to have a dimensionless quantity whose value is unaffected by changes in the unit used for the test score. The ratio between any two of the following three quantities, for example, could serve this purpose:

$$\begin{array}{ll} \text{observed-score variance} & \sigma_Y^2 \equiv \sigma^2(Y_{g*}), \\ \text{true-score variance} & \sigma_\zeta^2 \equiv \sigma^2(\zeta_*) \quad \text{or} \quad \sigma_T^2 \equiv \sigma^2(T_{g*}), \\ \text{error variance} & \sigma^2(\epsilon_{**}) \quad \text{or} \quad \sigma_E^2 \equiv \sigma^2(E_{g*}). \end{array}$$

A ratio that should probably be used more frequently is the *signal-to-noise ratio*  $\sigma_T^2/\sigma_E^2$  (see Section 5.12). This ratio meets the need mentioned at the end of the first paragraph, a need that cannot be met simply by a knowledge of  $\sigma_E^2$ .

The ratio most commonly used is the specific reliability coefficient  $\rho_{YT}^2$ , defined in Section 3.4 as the variance ratio

$$\rho_{YT}^2 = \sigma_T^2/\sigma_Y^2. \tag{9.1.1}$$

Equivalent definitions of the specific reliability coefficient (see Sections 3.2 and 3.3) include

$$\rho_{YT}^2 = \frac{1}{1 + \sigma_E^2/\sigma_T^2}, \quad (9.1.2)$$

$$\rho_{YT}^2 = 1 - \sigma_E^2/\sigma_Y^2, \quad (9.1.3)$$

$$\rho_{YT}^2 \equiv \rho^2(Y_{g*}, T_{g*}), \quad (9.1.4)$$

$$\rho_{YT}^2 = \rho(Y_{g*k}, Y_{g*l}), \quad (9.1.5)$$

where  $k$  and  $l$  denote parallel measurements. The last of these definitions (see Eq. 3.3.5b) is the only one that depends entirely on observable variables.

If  $\sigma_E = 3$  and  $\sigma_T = 15$ , we usually interpret the reliability coefficient  $\rho_{YT}^2 = 0.96$ , computed from (9.1.2), to mean that the observed scores correlate highly with true scores or, equivalently, that errors of measurement are small compared with true differences among examinees. If  $\sigma_E = 3$  and  $\sigma_T = 3$ , we usually interpret the reliability coefficient of 0.5 to mean that the observed score correlates poorly with true scores—that true-score differences among individual examinees have been blurred or obliterated by errors of measurement.

The reader should note that without further information, the reliability coefficient alone is of little value for describing a test as a measuring instrument. The reason is that a large reliability coefficient can often be obtained by administering the test to a sufficiently heterogeneous group of examinees (see Section 6.2). Thus reliability coefficients of 0.96 and 0.50 might both be found for a single test, depending on the group to which it is administered.

Because of this, the *Standards for Educational and Psychological Tests and Manuals* of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education (1966) urges that an observed-score variance always be reported along with each reliability coefficient. The fact is that at least two second-order moments or their equivalents (for example,  $\rho_{YT}^2$  and  $\sigma_Y^2$ , or  $\rho_{YT}^2$  and  $\sigma_E^2$ ) are necessary to describe the effectiveness of a test as a measuring instrument.

Sections 9.2, 9.4, 9.5, and 9.6 consider different methods for estimating the specific reliability of a test. Section 9.3 discusses the statistical properties of the ratio estimates in common use. Sections 9.7 and 9.8 cover generic reliability theory. Section 9.9 summarizes the use and interpretation of the coefficients discussed up to this point. Section 9.10 outlines some important but commonly ignored considerations about reliability, considerations that lead to some reservations about the suitability of many reliability coefficients used in common testing situations. The two final sections cover advanced material not necessary to the main development.

At first reading, the reader may wish to focus his attention on the theoretical principles that lead to variously defined reliability coefficients. He may leave the detailed methods for estimating these coefficients for later study.

## 9.2 Estimating the Specific Reliability Coefficient

In test theory, the conventional procedure for estimating such variance ratios as (9.1.1) is to replace numerator and denominator by unbiased sample estimates. (We shall consider the statistical properties of such an estimate of a variance ratio in Section 9.3.) Thus, when  $r$  replicate measurements on a single test are available, we can use (9.1.1) to write an estimate for the specific reliability coefficient:

$$\hat{\rho}_{YT}^2 \equiv \frac{\hat{\sigma}^2(T_{g*})}{\hat{\sigma}^2(Y_{g**})}. \quad (9.2.1)$$

For the numerator  $\hat{\sigma}^2(T_{g*})$ , we can take the unbiased estimate of the specific true-score variance given by (7.5.1). But what unbiased estimate of  $\sigma^2(Y_{g**}) = \sigma^2(Y_{g*})$  shall we use for the denominator? The symbol  $\hat{\sigma}^2(Y_{g**})$  has not been previously defined.

A commonly used unbiased estimate for  $\sigma^2(Y_{g**})$  is the within-replications mean square for test  $g$ :

$$\frac{1}{r} \frac{N}{N-1} \sum_{k=1}^r s^2(y_{g*k}). \quad (9.2.2)$$

However, this estimate does not use the information contained in such quantities as  $y_{g,l} - y_{g,k}$  (see the discussion following Eq. 8.8.8). In view of this, the estimate of  $\sigma^2(Y_{g**})$  that we choose is

$$\hat{\sigma}^2(Y_{g**}) \equiv \hat{\sigma}^2(T_{g*}) + \hat{\sigma}^2(E_{g*}), \quad (9.2.3)$$

the terms on the right being defined by (7.5.1) and (7.3.17). As noted in connection with (8.8.8), this estimate utilizes such differences as  $y_{g,l} - y_{g,k}$ . Thus *the estimate used here for the specific reliability coefficient is*

$$\hat{\rho}_{YT}^2 \equiv \frac{\hat{\sigma}^2(T_{g*})}{\hat{\sigma}^2(T_{g*}) + \hat{\sigma}^2(E_{g*})}, \quad (9.2.4)$$

where  $\hat{\sigma}^2(E_{g*})$  is the estimate given by (7.3.17) for the group specific error variance. More explicitly,

$$\hat{\rho}_{YT}^2 = \frac{Ns^2(y_{g**}) - \frac{1}{r-1} \frac{N-1}{N} \sum_{a=1}^N s^2(y_{ga*})}{Ns^2(y_{g**}) + \frac{N-1}{N} \sum_a s^2(y_{ga*})}. \quad (9.2.5)$$

The interested reader may wish to refer to the computations shown in Table 7.3.1, where  $\hat{\sigma}^2(Y_{g**}) = 150.57$  and  $\hat{\rho}_{YT}^2 = 0.9342$  were computed for the data presented in that table. For these same data, the mean square of (9.2.2) is 150.22; if this were divided into  $\hat{\sigma}^2(T_{g*})$ , the resulting estimate of reliability would be 0.9364.

For the case where  $Y_{g*}$  is normally distributed, Kristof has derived maximum likelihood estimates for  $\rho_{YT}^2$ , obtained the sampling distributions of the estimates, and shown how confidence intervals may be calculated. These developments are deferred until Section 9.5.

When rigorously parallel measurements are not available, that is, when  $r = 1$ , neither  $\hat{\sigma}^2(T_{g*})$  of (9.2.3) and (7.5.1) nor  $\hat{\sigma}^2(Y_{g**})$  of (9.2.3) can be obtained. Since there are no replications available to complicate matters, it seems natural to employ the usual unbiased estimate of observed-score variance

$$\hat{\sigma}^2(Y_{g*}) \equiv \frac{1}{N-1} \sum_{a=1}^N (y_{ga} - y_{g*})^2 = \frac{N}{N-1} s^2(y_{g*}), \quad (9.2.6)$$

instead of (9.2.3). The reader should note that although  $\sigma^2(Y_g) \equiv \sigma^2(Y_{g**})$ , the sample estimates denoted here by  $\hat{\sigma}^2(Y_{g*})$  and  $\hat{\sigma}^2(Y_{g**})$  differ because they are based on different kinds of data.

The numerator of (9.2.1) cannot be estimated when  $r = 1$ . However, we do have available the generous estimate  $\tilde{\sigma}_e^2$  of (7.7.1). We can obtain a useful approximation to  $\rho_{YT}^2$  by substituting  $\tilde{\sigma}_e^2$  and  $\hat{\sigma}^2(Y_{g*})$  into reliability formula (9.1.3):

$$\tilde{\rho}_{YT}^2 \equiv 1 - \frac{\tilde{\sigma}_e^2}{\hat{\sigma}^2(Y_{g*})}. \quad (9.2.7)$$

More explicitly, from (7.7.3) and (9.2.6),

$$\tilde{\rho}_{YT}^2 \equiv 1 - \frac{\sum_{g=1}^n s^2(y_{g*}) - ns^2(y_{**})}{(n-1)s^2(y_{g*})}. \quad (9.2.8)$$

Equation (9.2.8) provides a consistent\* estimate of a lower bound to the specific reliability coefficient of test  $g$  for situations where only nominally parallel measurements are available. The value of  $\tilde{\rho}_{YT}^2$  computed from this formula for the data in Table 8.4.1 is  $\tilde{\rho}_{YT}^2 = 0.947$  for test 1 and  $\tilde{\rho}_{YT}^2 = 0.928$  for test 2.

### 9.3 Statistical Properties of an Estimated Variance Ratio

In (9.2.4), a ratio of unbiased variance estimates was used to estimate the corresponding population variance ratio, i.e., to estimate the specific reliability coefficient. A similar statement holds for (9.2.7). The statistical properties of such a ratio of unbiased estimates deserve some attention here. Olkin and Pratt (1958) give detailed results for the case where multivariate normality is assumed. Here, however, we shall make no distributional assumptions.

---

\* To say that an estimate is consistent means that with as high probability as desired, the difference between the estimated and the true value of the parameter can be made as small as desired by making the sample size sufficiently large.

We can obtain the expected value of  $\hat{\rho}_{YT}^2$  over random sampling of people and of replications by expanding (9.2.4) in a Taylor series. To keep the notation in hand, we shall temporarily write  $\sigma_Y$  and  $\sigma_T$  instead of  $\sigma(Y_{g*})$  and  $\sigma(T_{g*})$ . The series expansion is

$$\begin{aligned}\hat{\rho}_{YT}^2 &\equiv \frac{\hat{\sigma}_T^2}{\hat{\sigma}_Y^2} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_Y^2 - \sigma_Y^2 + \sigma_Y^2} = \frac{\hat{\sigma}_T^2}{\sigma_Y^2} \left( \frac{\hat{\sigma}_Y^2 - \sigma_Y^2}{\sigma_Y^2} + 1 \right)^{-1} \\ &= \frac{\hat{\sigma}_T^2}{\sigma_Y^2} \left[ 1 - \frac{\hat{\sigma}_Y^2 - \sigma_Y^2}{\sigma_Y^2} + \frac{(\hat{\sigma}_Y^2 - \sigma_Y^2)^2}{\sigma_Y^4} - \frac{(\hat{\sigma}_Y^2 - \sigma_Y^2)^3}{\sigma_Y^6} + \dots \right], \quad (9.3.1)\end{aligned}$$

provided that  $|\hat{\sigma}_Y^2 - \sigma_Y^2| < \sigma_Y^2$ . Now  $\hat{\sigma}_Y^2$  is a *consistent estimate* of  $\sigma_Y^2$ . This means that with as high probability as desired, the difference  $|\hat{\sigma}_Y^2 - \sigma_Y^2|$  can be made as small as desired by taking  $N$  sufficiently large. Thus (9.3.1) can be used in large-sample work; furthermore  $(\hat{\sigma}_Y^2 - \sigma_Y^2)^2$  and subsequent terms can be neglected.

Taking expectations of (9.3.1), neglecting terms after the second, and using the symbol  $\doteq$  to indicate a large-sample approximation, we have

$$\mathcal{E}_a \mathcal{E}_k \hat{\rho}_{YT}^2 \doteq \frac{\sigma_T^2}{\sigma_Y^2} - \frac{1}{\sigma_Y^4} \mathcal{E}_a \mathcal{E}_k \hat{\sigma}_T^2 (\hat{\sigma}_Y^2 - \sigma_Y^2) \doteq \rho_{YT}^2 - \frac{1}{\sigma_Y^4} \text{Cov}_{ak}(\hat{\sigma}_T^2, \hat{\sigma}_Y^2). \quad (9.3.2)$$

The covariance in (9.3.2) is usually positive, so the estimate  $\hat{\rho}_{YT}^2$  is usually negatively biased. If  $\hat{\sigma}_Y^2 = \hat{\sigma}_T^2 + \hat{\sigma}_E^2$  and if  $T$  and  $E$  are distributed independently, then the covariance is always positive and  $\hat{\rho}_{YT}^2$  is always negatively biased. The covariance varies inversely as  $N$ ; consequently the bias in the estimate  $\hat{\rho}_{YT}^2$  approaches zero as  $N$  becomes large.

We could estimate and correct for the covariance in (9.3.2) and thus obtain an estimate of  $\rho_{YT}^2$  that is unbiased up to terms of order  $1/N^2$ . If we avoid special distributional assumptions, the correction term would involve fourth-order moments. In practical work, the sampling errors in estimating these moments might cause an unacceptable increase in the sampling error of the estimate of  $\rho_{YT}^2$ . If we do wish to reduce the bias in estimating  $\rho_{YT}^2$ , the best practical procedure may be Tukey's "jackknife" (Brillinger, 1964, Miller, 1964).

There is always a nonzero probability that  $\hat{\sigma}^2(T_{g*})$  will be negative (see Section 7.5); for discrete  $y_{ga}$  there is always a nonzero probability that  $\hat{\sigma}^2(Y_{g*})$  may be zero. We see that  $\hat{\rho}_{YT}^2$  may take on negative values and may even be negatively infinite. However, in practical testing work with aptitude and ability tests, reliability is normally above 0.5 and sample sizes are large enough so that negative sample estimates of reliability are very rare.

A theorem contributed by Slutsky (Cramér, 1946, p. 255) states that a ratio of consistent estimates is itself a consistent estimate of the corresponding ratio of parameters. It is clear from this or from (9.3.2) that  $\hat{\rho}_{YT}^2$  is a consistent estimate of  $\rho_{YT}^2$ . Statements similar to the foregoing apply also to similarly

constructed ratio estimates discussed in subsequent sections. A variance ratio estimate of a reliability coefficient is ordinarily a biased estimate, but at least it is a consistent estimate.

#### 9.4 Specific Reliability Theory for Composite Tests

A single mental test is very often a composite of separately scored parts. These may be either single test questions or entire (sub)tests. For convenience, we shall use the general term *test item* and the dummy indexes  $g$ ,  $h$ , and  $i$  to refer either to single test questions or to entire (sub)tests. As in Chapter 4, we shall use  $Y_1, Y_2, \dots, Y_g, \dots, Y_n$  to denote the scores on the  $n$  items, and  $X$  to denote the score on the total or *composite* test:

$$X \equiv \sum_{g=1}^n Y_g. \quad (9.4.1)$$

*If replicate measurements are available for a composite test, its specific reliability can be estimated by using (9.2.4), as with any other test, without breaking the composite down into subtests or items.* If replicate measurements are not available, then inferences about reliability can be obtained from item data, provided that the total test is not too heterogeneous. The present section presents a method for doing this.

As pointed out in Section 4.3, the specific error score  $E$  of a composite test is the sum of the specific error scores of the subtests or items:

$$E = \sum_{g=1}^n E_g.$$

Since the errors of measurement are uncorrelated, the group specific error variance for a composite test, as in (4.3.9), is

$$\sigma_E^2 = \sum_{g=1}^n \sigma^2(E_{g*}).$$

If we compare this equation with (7.6.8), which reads

$$\sigma^2(E_{**}) = \mathcal{E}_g \sigma^2(E_{g*}),$$

we see that  $\sigma_E^2$ , the group specific error variance for a composite test, is the same as  $n\sigma^2(E_{**})$  in the special case where there are only  $G = n$  subtests or items in the population under consideration. It follows that  $n\tilde{\sigma}_e^2$ , the generous estimate multiplied by  $n$ , is an unbiased estimate, over random sampling of people, for an upper bound to the group specific error variance of a composite test:

$$\mathcal{E}_a n \tilde{\sigma}_e^2 \geq \sigma_E^2. \quad (9.4.2)$$

The generous estimate can be computed from an analysis of variance components like that in Table 7.7.1 except that “tests” are replaced by “items”.

By the reasoning we used to obtain (9.2.7), we may also obtain an approximation to the specific reliability coefficient of a composite test:

$$\hat{\alpha} \equiv 1 - \frac{n\hat{\sigma}_e^2}{\hat{\sigma}^2(X_*)}, \quad (9.4.3)$$

where

$$\hat{\sigma}^2(X_*) \equiv \frac{1}{N-1} \sum_{a=1}^N (x_a - \bar{x}_*)^2 \equiv \frac{N}{N-1} s_x^2. \quad (9.4.4)$$

More explicitly, from (9.4.3) and (7.7.3),

$$\hat{\alpha} = 1 - \frac{n \sum_{g=1}^n s^2(y_{g*}) - n^2 s^2(y_{**})}{(n-1)s_x^2}.$$

Now, by definition,  $x_a = ny_{a*}$ ; therefore  $n^2 s^2(y_{**}) = s_x^2$  and, finally,

$$\begin{aligned} \hat{\alpha} &= 1 - \frac{\text{interaction mean square}}{\text{among-persons mean square}} \\ &= \frac{n}{n-1} \left[ 1 - \frac{1}{s_x^2} \sum_{g=1}^n s^2(y_{g*}) \right]. \end{aligned} \quad (9.4.5)$$

This is a sample analog of the coefficient  $\alpha$ , which has been discussed in detail in Section 4.4.

*When no replicate measurements are available,  $\hat{\alpha}$  in (9.4.5) provides an approximation to the specific reliability of a composite measurement  $X = \sum_g Y_g$ . Since  $\hat{\sigma}_e^2$  tends to be an overestimate, it follows that in sufficiently large samples of examinees,  $\hat{\alpha}$  is an underestimate of the specific reliability coefficient. (In the limiting case they may be equal, but this does not occur in practice.)*

For dichotomous test items, Aoyama (1957) has given the standard error of (9.4.5) under (1) random sampling of examinees, and (2) random sampling of test items. The formulas are too complex to reproduce here.

## 9.5 Maximum Likelihood Estimation of Reliability for Normally Distributed Scores

The important results in this and the following section are due to Kristof (1963a), who treated the general case in which a composite test  $X$  can be split into  $n$  parallel subtests. Here we deal only with the special case where a composite test  $X$  is split into two parallel half-tests  $Y_1$  and  $Y_2$ . By (4.2.10),

$$\rho_{XX'} = \frac{2\rho_{12}}{1 + \rho_{12}}, \quad (9.5.1)$$

$\rho_{XX'}$  being the reliability of the entire test and  $\rho_{12} \equiv \rho(Y_{1*}, Y_{2*})$  being the correlation between the half-tests.

In this and in the following section only, we assume that  $Y_1$  and  $Y_2$  have a bivariate normal frequency distribution. Since the two half-tests are parallel, their distribution is

$$\begin{aligned} f(y_1, y_2) &\equiv \frac{1}{2\pi\sigma^2\sqrt{1-\rho_{12}^2}} \\ &\times \exp\left[-\frac{(y_1-\mu)^2 + (y_2-\mu)^2 - 2\rho_{12}(y_1-\mu)(y_2-\mu)}{2\sigma^2(1-\rho_{12}^2)}\right], \end{aligned} \quad (9.5.2)$$

where  $\mu$  and  $\sigma$  are the mean and variance of  $Y_1$  and also of  $Y_2$ . If observations  $y_{1a}$  and  $y_{2a}$  are obtained for a random sample of  $a = 1, 2, \dots, N$  examinees, the logarithm of the likelihood function is

$$\begin{aligned} \log\left[\prod_{a=1}^N f(y_{1a}, y_{2a})\right] &= -N \log 2\pi - 2N \log \sigma - \frac{N}{2} \log (1 - \rho_{12}^2) \\ &- \frac{\sum_{a=1}^N (y_{1a} - \mu)^2 + \sum_{a=1}^N (y_{2a} - \mu)^2 - 2\rho_{12} \sum_{a=1}^N (y_{1a} - \mu)(y_{2a} - \mu)}{2\sigma^2(1 - \rho_{12}^2)}. \end{aligned} \quad (9.5.3)$$

Differentiating this with respect to each of the parameters  $\mu$ ,  $\sigma$ , and  $\rho_{12}$  and setting each derivative equal to zero yields three likelihood equations, whose solutions (Jackson and Ferguson, 1941, Eq. 85) are the maximum likelihood estimates, denoted here by ( $\hat{\cdot}$ ):

$$\begin{aligned} \hat{\mu} &\equiv \frac{\sum_{a=1}^N y_{1a} + \sum_{a=1}^N y_{2a}}{2N}, \\ \hat{\sigma}^2 &\equiv \frac{\sum_{a=1}^N y_{1a}^2 + \sum_{a=1}^N y_{2a}^2}{2N} - \hat{\mu}^2, \\ \hat{\rho}_{12} &\equiv \frac{1}{\hat{\sigma}^2} \left( \frac{1}{N} \sum_{a=1}^N y_{1a}y_{2a} - \hat{\mu}^2 \right). \end{aligned} \quad (9.5.4)$$

From (9.5.1) and (9.5.4), we find the maximum likelihood estimate of  $\rho_{XX'}$  to be

$$\begin{aligned} \hat{\rho}_{XX'} &\equiv \frac{2\hat{\rho}_{12}}{1 + \hat{\rho}_{12}} = 1 - \frac{1 - \hat{\rho}_{12}}{1 + \hat{\rho}_{12}} = 1 - \frac{2N\hat{\sigma}^2 - 2 \sum_{a=1}^N y_{1a}y_{2a} + 2N\hat{\mu}^2}{2N\hat{\sigma}^2 + 2 \sum_{a=1}^N y_{1a}y_{2a} - 2N\hat{\mu}^2} \\ &= 1 - \frac{\sum_{a=1}^N y_{1a}^2 + \sum_{a=1}^N y_{2a}^2 - 2 \sum_{a=1}^N y_{1a}y_{2a}}{\sum_{a=1}^N y_{1a}^2 - Ny_{1.}^2 + \sum_{a=1}^N y_{2a}^2 - Ny_{2.}^2 + 2 \sum_{a=1}^N y_{1a}y_{2a} - 2Ny_{1.}y_{2.}} \\ &= 1 - \frac{1}{Ns_x^2} \sum d_a^2, \end{aligned}$$

where  $d_a = y_{2a} - y_{1a}$ , and

$$s_x^2 \equiv \frac{1}{N} \sum_{a=1}^N (x_a - \bar{x})^2$$

is the sample variance of the total test score. Thus, if a test  $X$  can be split into two parallel halves, the scores on the halves being bivariate-normally distributed, then the maximum likelihood estimate of the reliability of test  $X$  is

$$\hat{\rho}_{XX'} \equiv 1 - \frac{1}{Ns_x^2} \sum_{a=1}^N d_a^2. \quad (9.5.5)$$

If the maximum likelihood estimator  $s_x^2$  in this formula is replaced by the unbiased estimator  $Ns_x^2/(N - 1)$ , it can be shown that the resulting estimate of reliability is the same as that obtained for  $r = 2$  by stepping up (9.2.5) by the Spearman-Brown formula (4.2.10). The reliability of the composite score  $X$  for the two tests in Table 7.3.1 is estimated from (9.5.5) to be

$$\hat{\rho}_{XX'} = 1 - \frac{198}{10(524.24)} = 0.9622.$$

The reliability obtained for the same data by using (9.2.5) and then stepping the result up to double length is

$$\hat{\rho}_{X\Gamma}^2 = \frac{2(0.9342)}{1 + 0.9342} = 0.9660.$$

## 9.6 The Frequency Distribution of the Estimated Reliability

In this section, we extend the discussion of the preceding section while retaining the same assumptions about the random variables  $Y_1$  and  $Y_2$ . It is easily verified that the covariance between the random variables  $X \equiv Y_1 + Y_2$  and  $D \equiv Y_2 - Y_1$  is zero:

$$\sigma(Y_2 + Y_1, Y_2 - Y_1) = \sigma_2^2 - \sigma_1^2 = 0. \quad (9.6.1)$$

It is well known that any weighted sum of multivariate normal variables is itself normally distributed. Thus we see that  $X$  is normally distributed with variance

$$\sigma_X^2 = \sigma^2(Y_1 + Y_2) = \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho_{12} = 2\sigma^2(1 + \rho_{12}), \quad (9.6.2)$$

and that  $D \equiv Y_2 - Y_1$  is independently normally distributed with variance

$$\sigma_D^2 = 2\sigma^2(1 - \rho_{12}). \quad (9.6.3)$$

It follows (Freeman, 1963, Chapter 23) that  $Ns_x^2/2\sigma^2(1 + \rho_{12})$  has a chi-square distribution with  $N - 1$  degrees of freedom, and also, since

$$\mathcal{E}D = \mathcal{E}Y_2 - \mathcal{E}Y_1 = 0,$$

it follows that

$$\frac{1}{2\sigma^2(1 - \rho_{12})} \sum_a^N d_a^2$$

has an *independent* chi-square distribution with  $N$  degrees of freedom. If each chi-square variable is divided by the corresponding degrees of freedom, the ratio of the two quotients will have an  $F$ -distribution (Freeman, 1963, Section 23.10). Thus

$$F_{N,N-1} \equiv \frac{N-1}{N} \frac{\sum_a^N d_a^2}{Ns_x^2} \frac{1 + \rho_{12}}{1 - \rho_{12}} = \frac{N-1}{N} \frac{1 - \hat{\rho}_{XX'}}{1 - \rho_{XX'}} \quad (9.6.4)$$

has an  $F$ -distribution with  $N$  and  $N - 1$  degrees of freedom. Equally

$$F_{N-1,N} \equiv \frac{N}{N-1} \frac{1 - \rho_{XX'}}{1 - \hat{\rho}_{XX'}} \quad (9.6.5)$$

has an  $F$ -distribution with  $N - 1$  and  $N$  degrees of freedom.

The distribution of  $\hat{\rho}_{XX'}$  depends on the unknown parameter  $\rho_{XX'}$ . Since the distribution of  $F$  is well tabled, the probability that  $\hat{\rho}_{XX'}$  lies below any specified value can be deduced from these tables, with the help of (9.6.5), for any specified value of the parameter  $\rho_{XX'}$ . Thus the (small-sample) distribution of  $\hat{\rho}_{XX'}$  can be considered a known function of  $\rho_{XX'}$ .

To illustrate the application of this result, suppose that the maximum likelihood estimate of the reliability of test  $X$  has been computed from data on just  $N = 10$  examinees and found to be  $\hat{\rho}_{XX'} = 0.75$ . Is it implausible to assume that the true reliability is close to zero?

A 95% one-tailed confidence interval for  $\rho_{XX'}$  may be obtained as follows:

Let

$$\text{Prob } (F_{N-1,N} \leq F_{0.95}) = 0.95,$$

where  $F_{0.95} = 3.02$  is a tabled constant of the  $F$ -distribution. Since

$$\frac{N}{N-1} \frac{1 - \rho_{XX'}}{1 - \hat{\rho}_{XX'}} = \frac{40}{9} (1 - \rho_{XX'}),$$

then, by (9.6.5),

$$\text{Prob } [\frac{40}{9}(1 - \rho_{XX'}) \leq 3.02] = 0.95 \quad \text{or} \quad \text{Prob } (\rho_{XX'} \geq 0.3205) = 0.95.$$

Since the 95% one-tailed confidence interval for the test reliability goes down only to  $\rho_{XX'} = 0.32$ , it appears that the test has at least an appreciable degree of reliability. In this case, ten observations were enough to provide this much information about the test reliability coefficient.

Kristof used (9.6.4) and (9.6.5) to find that *under the assumptions of the preceding section, the reliability estimate defined by*

$$\hat{\rho}_{XX'}^* \equiv \frac{3}{N} + \frac{N-3}{N} \hat{\rho}_{XX'} \quad (9.6.6)$$

is unbiased over random sampling of examinees. This result contrasts with a result of Olkin and Pratt (1958) who found no simple unbiased estimate of the (not stepped-up) correlation between parallel variables. For the data in Table 7.3.1, we find Kristof's unbiased estimate to be

$$\hat{\rho}_{XX'}^* = \frac{3}{10} + \frac{7}{10}(0.9622) = 0.9735.$$

### 9.7 The Generic Reliability Coefficient

The generic reliability coefficient, which we shall denote by  $\rho_{Y\xi}^2$ , may be defined as

$$\rho_{Y\xi}^2 \equiv \frac{\sigma^2(\xi_*)}{\sigma^2(Y_{**})}, \quad (9.7.1)$$

the ratio of the variance over people of the generic true score  $\xi_*$  to the variance over people and tests of the observed score  $Y_{**}$ . The reader will remember that the generic true score is the expected score over nominally parallel tests (Section 8.1). As in Chapter 8, an infinite population of tests is assumed in all discussions of generic parameters and statistics unless otherwise specified.

Now the generic errors  $\epsilon_{**}$  are unbiased and uncorrelated with each other and with generic true score (Section 8.4) provided that all expectations are taken over all tests. It follows by the same reasoning used in Sections 3.2 and 3.3 that Equations (9.7.1), (9.7.2), (9.7.3), and (9.7.4) are equivalent definitions for the generic reliability coefficient:

$$\rho_{Y\xi}^2 \equiv \frac{1}{1 + \sigma^2(\epsilon_{**})/\sigma^2(\xi_*)}, \quad (9.7.2)$$

$$\rho_{Y\xi}^2 \equiv 1 - \frac{\sigma^2(\epsilon_{**})}{\sigma^2(Y_{**})}, \quad (9.7.3)$$

$$\rho_{Y\xi}^2 \equiv \rho^2(Y_{**}, \xi_*). \quad (9.7.4)$$

This coefficient is appropriate for use when the generic true score and the group generic error variance are the quantities of basic interest. *The generic reliability coefficient is appropriate when the data are unmatched, that is, when tests are assigned to examinees at random.*

The generic reliability coefficient  $\rho_{Y\xi}^2$  describes a whole genus of nominally parallel forms. No definition that involves  $\rho(Y_{g*}, Y_{h*})$ , as does (9.1.5), is of value in defining such a coefficient since  $\rho(Y_{g*}, Y_{h*})$  will normally be different for each different choice of  $g$  and  $h$ . The average of all such correlation coefficients does not have a simple mathematical relation to other quantities of interest.

*A consistent estimate of the generic reliability coefficient is given by*

$$\hat{\rho}_{Y\xi}^2 \equiv \frac{\hat{\sigma}^2(\xi_*)}{\hat{\sigma}^2(\xi_*) + \hat{\sigma}^2(\epsilon_{**})}, \quad (9.7.5)$$

the quantities on the right being the estimates for the generic true-score and

error variances given by (8.5.3) or (8.5.6), and (8.3.7). We justify (9.7.5) by the same reasoning used to justify (9.2.4). When substitutions are made from (8.5.6) and (8.3.7), this formula becomes the same (except for a typographical error) as that given by Cronbach *et al.* (1963, Eq. 34).

For  $n = 2$ , Eq. (9.7.5) can be written explicitly as

$$\hat{\rho}_{Y\zeta}^2 = \frac{s(y_{1*}, y_{2*})}{s(y_{1*}, y_{2*}) + [(N - 1)/2N^2] \sum_{a=1}^N d_a^2}. \quad (9.7.6)$$

If we compute this quantity for the data in Table 8.4.1, we find that

$$\hat{\rho}_{Y\zeta}^2 = \frac{126.92}{126.92 + 9(198)/200} = 0.934.$$

## 9.8 Generic Reliability for a Single Test

We pointed out in the preceding section that the generic reliability coefficient  $\rho_{Y\zeta}^2 \equiv \rho^2(Y_{**}, \zeta_*)$  defined there describes a whole genus of nominally parallel test forms. In the present section, we are concerned with the problem of describing a single test form  $g$ .

Suppose that an arithmetic teacher has compiled  $n = 4$  nominally parallel final examinations for an arithmetic course. If each student is randomly assigned one of these tests at the end of the course, the generic reliability of the overall measurement procedure, including the random selection of test forms, can be represented by  $\rho_{Y\zeta}^2 \equiv \rho^2(Y_{**}, \zeta_*)$ , the parameter treated in the preceding section. On the other hand, if the teacher has administered only test  $g$ , he will wish to know as much as possible about the properties of the measurements he has actually obtained. In this case, he needs a generic reliability coefficient describing the measurements obtained from the single test form  $g$ .

In Section 9.1, we listed five equivalent definitions for the specific reliability coefficient for test  $g$ . In Section 9.7, we listed four equivalent definitions for the generic reliability coefficient describing an entire genus of nominally parallel test forms. The equivalence we noted within each set of definitions resulted from the fact that in each situation the errors of measurement were uncorrelated with other variables. As shown in Section 8.4, we cannot assume this non-correlation for the generic errors of measurement on a given test  $g$ . In fact, we can readily show that no two of the definitions of reliability that were equivalent in Section 9.7 will be equivalent for describing a given test  $g$ .

For example, writing  $\sigma_g^2$  instead of  $\sigma^2(Y_{g*})$ , and  $\rho_{g\zeta}$  instead of  $\rho(Y_{g*}, \zeta_*)$ , we find that

$$\begin{aligned} \rho_{g\zeta}^2 &= \frac{\sigma_{g\zeta}^2}{\sigma_g^2 \sigma_\zeta^2} = \frac{\sigma^2(\zeta_* + \epsilon_{g*}, \zeta_*)}{\sigma_g^2 \sigma_\zeta^2} = \frac{[\sigma_\zeta^2 + \sigma(\epsilon_{g*}, \zeta)]^2}{\sigma_g^2 \sigma_\zeta^2} \\ &= \frac{\sigma_\zeta^2}{\sigma_g^2} \left[ 1 + \frac{\sigma(\epsilon_{g*})}{\sigma_\zeta} \rho_{\epsilon\zeta} \right]^2 = \frac{\sigma_\zeta^2}{\sigma_g^2} (1 + \beta_{\epsilon\zeta})^2, \end{aligned} \quad (9.8.1)$$

where  $\beta_{\epsilon\xi}$  is the regression coefficient of  $\epsilon_g$  on  $\xi$ , and  $\sigma_{\epsilon\xi}$  and  $\rho_{\epsilon\xi}$  are the corresponding covariance and correlation. Equation (9.8.1) shows that the reliability coefficient  $\rho_{g\xi}^2$  differs from the familiar variance ratio  $\sigma_\xi^2/\sigma_g^2$  by a factor of  $(1 + \beta_{\epsilon\xi})^2$ .

We can easily show further that when  $\epsilon_g$  and  $\xi$  are positively correlated,

$$\rho_{g\xi}^2 \geq 1 - \frac{\sigma^2(\epsilon_{g*})}{\sigma_g^2} \geq \frac{\sigma_\xi^2}{\sigma_g^2}. \quad (9.8.2)$$

When  $\epsilon_g$  and  $\xi$  are negatively correlated, the size relation of the three coefficients is not so simple. It is worth noting, however, that when  $\epsilon_g$  and  $\xi$  are negatively correlated,  $\sigma_\xi^2/\sigma_g^2$  may be greater than 1.

Cronbach *et al.* (1963, Sections 3 and 4) give formulas for estimating both  $\rho_{g\xi}^2$  and  $\mathcal{E}_g(\sigma_\xi^2/\sigma_g^2)$  (see also Cronbach, Ikeda, and Avner, 1964). Although they suggest estimates of the latter coefficient for practical use, we prefer for most purposes to estimate  $\rho_{g\xi}^2$  itself. One of the reasons for this preference is the fact that  $\sigma_\xi^2/\sigma_g^2$  may exceed 1.

The estimate of  $\rho_{g\xi}^2$  that suggests itself is

$$\hat{\rho}_{g\xi}^2 \equiv \frac{\hat{\sigma}_{g\xi}^2}{\hat{\sigma}_g^2 \hat{\sigma}_\xi^2}, \quad (9.8.3)$$

where

$$\hat{\sigma}_{g\xi}^2 \equiv \hat{\sigma}(Y_{g*}, \xi_*)$$

is given by (8.7.13) or (8.7.14), where  $\hat{\sigma}_g^2 \equiv \hat{\sigma}^2(Y_{g*})$  is given by (9.2.6), and where  $\hat{\sigma}_\xi^2$  is given either by (8.5.6) with  $n$  replaced by  $n'$  or by (8.7.12). Explicitly,

$$\rho_{g\xi}^2 \equiv \frac{(n' - 1) \left[ \sum_{h \neq g} s(y_{g*}, y_{h*}) \right]^2}{2n's^2(y_{g*}) \sum_{h > i} \sum s(y_{h*}, y_{i*})}, \quad (9.8.4)$$

where  $h, i \neq g$  and  $h, i = 1, 2, \dots, n'$ , as in Section 8.7. Equation (9.8.4) gives a consistent estimate for the generic reliability of a particular test form  $g$ .

As with (8.7.8), the examiner must administer at least  $n = 3$  test forms (two in addition to form  $g$ ) to each student if he is to be able to use (9.8.4). If, as is usual, all the nominally parallel test forms available are much alike,  $n = 3$  will be adequate; if, on the other hand, the test forms are substantially different,  $n = 3$  will not be adequate. It is not difficult, however, to construct a data-gathering design, along lines we shall consider in Section 11.13, such that the examiner can use (9.8.4) to reliably estimate  $\rho_{g\xi}^2$  from data in which no one examinee takes more than two of the  $n > 3$  nominally parallel test forms. Detailed formulas will not be given here.

### 9.9 Use and Interpretation of Reliability Coefficients

If rigorously parallel measurements are available,

$$\hat{\rho}_{YT}^2 \equiv \frac{\hat{\sigma}^2(T_{g*})}{\hat{\sigma}^2(T_{g*}) + \hat{\sigma}^2(E_{g*})}$$

provides a satisfactory estimate of the specific reliability coefficient for a single measurement (see Eq. 9.2.4). A more explicit but more cumbersome formula is given as (9.2.5). The interpretation of the coefficient may follow any of the equivalent definitions (9.1.1) through (9.1.5). If normality can be assumed, then (9.6.6) has many advantages over (9.2.4). It is unbiased, and its sampling distribution is known.

If nominally but not rigorously parallel measurements are available, then the specific reliability coefficient of a single test can be approximated by (9.2.7):

$$\tilde{\rho}_{YT}^2 \equiv 1 - \frac{\tilde{\sigma}_e^2}{\tilde{\sigma}^2(Y_{g*})}.$$

A more explicit formula is given as (9.2.8). When the  $n$  available tests are not rigorously parallel, (9.2.7) will necessarily give an underestimate, provided that the number of examinees is sufficiently large.

The estimate (9.4.5) of coefficient  $\alpha$ ,

$$\hat{\alpha} \equiv \frac{n}{n-1} \left[ 1 - \frac{1}{s^2(x_*)} \sum_{g=1}^n s^2(y_{g*}) \right],$$

gives an approximation for the specific reliability of the composite score  $X_* \equiv \sum_g Y_{g*}$ . If the number of examinees is sufficiently large,  $\hat{\alpha}$  will be an underestimate—a slight underestimate if the test is long.

So-called parallel tests are seldom, if ever, strictly parallel. It is sometimes argued that parallel measurements may be obtained by administering the same test twice. The effect of memory, however, usually prevents the two sets of measurements from being strictly parallel. There are also other good reasons (see Section 8.1) for being interested in generic rather than specific reliability coefficients. This point is well expressed in American Psychological Association *et al.* (1966), *Standards for Educational and Psychological Tests and Manuals*:

Aside from practical limitations, retesting is not a theoretically desirable method of determining a reliability coefficient if, as usual, the items that constitute the test are only one of many sets (actual or hypothetical) that might equally well have been used to measure the particular ability or trait. Thus, there is ordinarily no reason to suppose that *one* set of (say) 50 vocabulary items is especially superior (or inferior) to another comparable (equivalent) set of 50. In this case it appears desirable to determine not only the degree of response variation by the subject from one occasion to the next (as is accomplished by the retest method), but also the extent of sampling fluctuation involved in selecting a given set of 50 items. These two objectives are accomplished

**Table 9.9.1**  
Estimates of reliability

Type of reliability estimated	Type of measurements required	Biased estimate?	Distributional assumptions	Symbol	Equation	Formula
Specific	Parallel	Biased	None	$\hat{\rho}_{YT}^2$	(9.2.5)	$\frac{\hat{\sigma}^2(T_{\theta^*})}{\hat{\sigma}^2(T_{\theta^*}) + \hat{\sigma}^2(E_{\theta^*})}$
Specific	Parallel	Unbiased	Normality	$\hat{\rho}_{XX'}^*$	(9.6.6)	$\frac{3}{N} + \frac{N-3}{N} \hat{\rho}_{XX'}$
Specific	Nominally parallel	(Estimates a lower bound)	None	$\tilde{\rho}_{YT}^2$	(9.2.8)	$1 - \frac{\tilde{\sigma}_e^2}{\hat{\sigma}^2(Y_{\theta^*})}$
Specific	Composite		None	$\hat{\alpha}$	(9.4.5)	$\frac{n}{n-1} \left[ 1 - \frac{\sum_{i=1}^n s^2(y_{\theta^*})}{s^2(x_i)} \right]$
Generic	Nominally parallel	Biased	None	$\hat{\rho}_{Y\xi}^2$	(9.7.5)	$\frac{\hat{\sigma}^2(\xi^*)}{\hat{\sigma}^2(\xi^*) + \hat{\sigma}^2(\epsilon^{**})}$
Generic	Nominally parallel	Biased	None	$\hat{\rho}_{\theta\xi}^2$	(9.8.4)	$\frac{\hat{\sigma}^2(Y_{\theta^*}, \xi^*)}{\hat{\sigma}^2(Y_{\theta^*}) \hat{\sigma}^2(\xi^*)}$

most commonly by correlating scores on the original set of 50 items with scores by the same subjects on an independent but similar set of 50 items—an ‘alternate form’ of the original 50. If the effect of content-sampling *alone* is sought (without the effects of response variability by the subject), or if it is not practical to undertake testing on two different occasions, a test of 100 items may be administered. Then the test may be divided into two sets of 50 odd-numbered items and 50 even-numbered items . . .\*

This line of reasoning leads to the use of a generic reliability coefficient. If the examiner wants a reliability coefficient that describes a whole genus of nominally parallel test forms, he can use the generic reliability coefficient  $\rho_{Y\zeta}^2$ . This is appropriate in a situation where tests are assigned to examinees at random. The estimate for the generic reliability of the whole genus is given as (9.7.5):

$$\hat{\rho}_{Y\zeta}^2 \equiv \frac{\hat{\sigma}^2(\zeta_*)}{\hat{\sigma}^2(\zeta_*) + \hat{\sigma}^2(\epsilon_{**})}.$$

When  $n = 2$ , (9.7.6) is an explicit formula for this estimate.

If, as is usual, the examiner wants to describe the measurement properties of a particular nominally parallel test form  $g$  in relation to the generic true score, he will use (9.8.3) to estimate the generic reliability:

$$\hat{\rho}_{g\zeta}^2 \equiv \hat{\rho}^2(Y_{g*}, \zeta_*) = \hat{\sigma}^2(Y_{g*}, \zeta_*) / \hat{\sigma}^2(Y_{g*})\hat{\sigma}^2(\zeta_*).$$

An explicit formula is given as (9.8.4). At least three nominally parallel sets of measurements, including those on test  $g$ , are needed to get a satisfactory estimate for  $\rho_{g\zeta}^2$ .

Suppose form A of a test has a higher specific reliability than form B, but form B has a higher generic reliability. Which form should be preferred? This situation can arise only if the two forms measure somewhat different things. Then, however, the choice of forms should not be made solely on the basis of reliability. The choice will depend on which true score we wish to measure.

A summary of reliability estimates is given in Table 9.9.1. The most common practical use of both specific and generic reliability coefficients is the same: to describe the correlation between the observed score and the appropriate true score. If this correlation is too low, the examinee’s observed score is not too helpful in reaching decisions that ought to depend on his true score. Either the specific or the generic reliability coefficient can be used to obtain linear regression estimates of the corresponding true score, if required. The specific reliability coefficient can be used to correct a validity coefficient for attenuation; the generic reliability coefficient by itself usually cannot, except under special assumptions. In the next section, we shall point out a fact that limits the value of both types of coefficients in most common applications.

---

\* From *Standards for educational and psychological tests and manuals*. Copyright 1966, American Psychological Association. Used by permission.

All the results given in Table 9.9.1 agree with formulas obtained by Cronbach *et al.* (1963). Theirs is the basic paper in which a number of the ideas presented here were developed and presented. They use the term *generalizability* where we retain the older term *reliability*. As explained in Chapter 8, we have not given generic formulas in terms of parameters estimated from unmatched data. These and other formulas are given by Cronbach, by Gleser, and by Rajaratnam in papers listed in the references at the end of this chapter.

The American Psychological Association (1966) standards points out that inconsistency among observations may arise from variations in (1) the subject's response, (2) test content, and (3) test administration, as well as from scoring errors and other sources. It continues:

*The estimation of clearly labeled components of error variance is the most informative outcome of a reliability study*, both for the test developer wishing to improve the reliability of his instrument and for the user desiring to interpret test scores with maximum understanding. The analysis of error variance calls for the use of an appropriate *experimental design*. There are many different multivariate designs that can be used in reliability studies; the choice of design for studying the particular test is to be determined by its intended interpretation and by the practical limitations upon experimentation.\*

The examiner must design an appropriate set of test administrations, one that allows him to isolate and estimate the desired sources of error. Given data so obtained, an extension of the variance components analysis approach of Chapters 7 and 8 will often produce the desired estimates; see Gleser, Cronbach, and Rajaratnam (1965), Collier, Baker, and Mandeville (1967), Cureton (1965), Medley and Mitzel (1963), Stanley (1962, 1961, 1955), Burt (1955), and Lindquist (1953). Other methods for analyzing data involving trends over time are discussed in C. C. Anderson (1958), T. W. Anderson (1963), Bock and Haggard (in press), Bock (1963), Church (1966), Cole and Grizzle (1966), Coleman (1964), Danford, Hughes, and McNee (1960), Fleiss (1966), Gaito and Wiley (1963), Garside (1958), Geisser (1959), Glaser (1952), Goldfarb (1960), Hoffman (1963), Holtzman (1963), Horn and Little (1966), Howard (1964), Lindquist (1947), McHugh and Wall (1962), Potthoff and Roy (1964), Rao (1958, 1965), Schaie (1965), Thorndike (1966), Tucker (1966), Vinsonhaler and Meredith (1966), and Wiggins (1965). See also Norman (1967).

## 9.10 The Reliability of Ordinal Measurements

In most work, reliability theory has been based on certain tacit and commonly ignored assumptions about the way the test scores are used and interpreted.

---

\* From *Standards for educational and psychological tests and manuals*. Copyright 1966, American Psychological Association. Used by permission.

The fact is that these assumptions do not hold strictly for many practical testing situations (see Buros, 1963).

Suppose an examiner is to hire or admit to college the top 20 or 80% of examinees tested. Or suppose he is to give a scholarship to the highest-scoring examinee or to the twenty highest-scoring examinees. In such cases, the examiner is using only the *ordinal* properties (the rank order) of the test scores. For his purposes, therefore, it would be inappropriate (except as a practical approximation) to describe the relation of observed score to true score by some parameter that changes unpredictably when the scores are monotonically but nonlinearly transformed. Neither the specific nor the generic error of measurement varies in any predictable way under such transformations; thus both would seem inappropriate for his purposes.

This and similar considerations lead us to the formulation of the following guiding principle:

*If there is a certain class of transformations on the observed scores that do not change the meaning and the use of the scores in a certain testing situation, then any parameter used to describe the relation of observed score to true score in this situation should be invariant under this class of transformations.*

Although the examiner is usually interested in a "true score" of some kind, it does not follow that he is always interested in the corresponding "error", defined as the difference between true and observed score. If the examiner is making use only of the ordinal properties of the score scale, he should be interested in the relationship between the rank order of the examinee's observed score and the rank order of his true score, rather than in the usual specific or generic error of measurement. *Most ordinary practical uses of a specific or a generic reliability coefficient violate this principle.*

If the rank correlation coefficient between observed and true scores (i.e., the product moment correlation between rankings on the two variables) could be estimated, it would be invariant under monotonic transformation of either variable and thus would satisfy the guiding principle. Unfortunately, however, it may be difficult or impossible to estimate such a coefficient for the usual type of true score without making strong distributional assumptions. The rank order of the usual true scores (expected observed scores) is ordinarily *not* the same as the expectation of the ranks of the observed scores.

One way around this problem is to treat each observed rank as an observed score, defining specific or generic true score as expected rank over repeated measurements. If this is done, the usual formulas for specific or generic error variance and reliability can be applied directly to the ranks. For example, the sample rank correlation between the parallel forms in Table 7.3.1 is 0.915. This is an estimate of specific reliability, provided that (1) only the rank order of the test scores is used in making decisions about examinees, and (2) true score is defined as the expectation of an examinee's rank.

### 9.11 Use of Factor Loadings as Reliability Coefficients\*

Consider a population of nominally parallel tests  $g = 1, 2, \dots$ , and suppose that standardized scores  $\mathcal{Y}_{g*}$  are computed on each test separately by the linear transformation

$$\mathcal{Y}_{g*} = \frac{Y_{g*} - \mathcal{E}_a(Y_{ga})}{\sigma(Y_{g*})}.$$

Note that the generic true score  $\mathcal{E}_g(Y_{g*})$  usually will not correlate perfectly with the generic true score  $\mathcal{E}_g(\mathcal{Y}_{g*})$ . Which generic true score is to be preferred? This dilemma suggests that we improve the definition of generic true score for situations where the interpretation of observed scores on a fixed test  $g$  is unaffected by any positive linear transformation applied to all test  $g$  scores.

If the test intercorrelations have just one common factor (see Section 16.7), a good procedure would seem to be to use the common factor in place of  $\zeta$ . Less restrictively, if the tests depend only on a one-dimensional latent trait  $\theta$ , then we can use the latent trait in place of  $\zeta$ ; the reader may consult Chapter 16, and especially Section 16.11, on this point. In either case, the square of the correlation between  $Y_{g*}$  and  $\theta$  can then be used as a reliability coefficient for test  $g$ .

If  $\theta$  is defined in accordance with modern methods of (common-) factor analysis (see, e.g., Lawley and Maxwell, 1963, Chapter 2) and not by principal components analysis, then  $\theta$  will not be affected by any linear transformation of one or more of the test scores. In the type of situation described, only such an invariant reliability coefficient can satisfy the guiding principle set forth in Section 9.10.

If the tests have more than one common factor or more than one latent trait, the correlation of  $Y_{g*}$  with each will be of interest. The reader may refer to LaForge (1965) and Rozeboom (1966).

### 9.12 Estimating Reliability Without Using Parallel Forms†

Suppose the available forms of a test are of unequal length, or of different degrees of difficulty, or of different specific reliability. The examiner will not wish to consider these forms as nominally parallel, particularly if they are of unequal length. If the examiner wishes, he may, under conditions to be specified, use a linear structural relations model (Kendall and Stuart, 1961, Chapter 29) or an equivalent factor analysis model to estimate the reliability of any particular test form. [Madansky (1964), without concerning himself with test reliability coefficients, describes some important relationships between the two specialized areas of factor analysis and linear structural relations.]

---

\* This section assumes some familiarity with factor analysis. It may be omitted without loss of continuity.

† This section treats a topic that may be omitted without loss of continuity.

Every typical problem in linear structural relations involves two or more unobservable variables that are linearly related to each other. For example, the true scores  $\xi$  and  $\eta$  on two tests  $X$  and  $Y$  might have the linear relationship

$$\eta = A\xi + B,$$

where  $A$  and  $B$  are parameters to be determined. Now  $\xi$  and  $\eta$  cannot be observed without errors of measurement  $E_x$  and  $E_y$ . The problem is to infer the parameters  $A$  and  $B$  of the linear relationships and also, in this example, some parameters (the variances) of the distributions of the errors of measurement, from fallible observations such as  $x = \xi + e_x$  and  $y = \eta + e_y$ . If observations were made directly on  $\xi$  and  $\eta$ , inferring  $A$  and  $B$  would be a simple regression problem. Since only the fallible measurements  $X$  and  $Y$  are available, however, we have a problem in linear structural relations.

Consider three test forms  $Y_1$ ,  $Y_2$ , and  $Y_3$ , with true scores  $T_1$ ,  $T_2$ , and  $T_3$  and error scores  $E_1$ ,  $E_2$ , and  $E_3$ . For examinee  $a$ ,

$$y_{1a} = \tau_{1a} + e_{1a}, \quad y_{2a} = \tau_{2a} + e_{2a}, \quad y_{3a} = \tau_{3a} + e_{3a}. \quad (9.12.1)$$

The assumption of parallelism would imply that  $T_1 \equiv T_2 \equiv T_3$  and that  $\sigma(E_{1*}) = \sigma(E_{2*}) = \sigma(E_{3*})$ , as in Sections 7.2, 7.3, and 9.2. Suppose this is too strong a set of assumptions for a certain set of data. We may still estimate error variances, true-score variances, and reliability coefficients if we can make the weak assumption that the random variables  $T_1$ ,  $T_2$ , and  $T_3$  are linearly related.

These true scores cannot be called either specific or generic. They are defined only through the assumptions made about them; nothing will be said about the existence either of parallel forms or of nominally parallel forms.

We shall denote the variances of  $T_1$ ,  $T_2$ , and  $T_3$  by  $\sigma(T_{g*})$ , and their covariances by  $\sigma(T_{g*}, T_{h*})$ ,  $g \neq h$ ,  $g, h = 1, 2, 3$ . Without loss of generality, the assumption that  $T_1$ ,  $T_2$ , and  $T_3$  are linearly related may be written

$$T_1 = A_1 T_0 + B_1, \quad T_2 = A_2 T_0 + B_2, \quad T_3 = A_3 T_0 + B_3, \quad (9.12.2)$$

where the random variable  $T_0$  is chosen to have mean  $\mu_T = 0$  and variance  $\sigma_T^2 = 1$  in the population of examinees;  $T_0$  is a new variable defined by Eq. (9.12.2). We quickly find from (9.12.2) that

$$A_g = \sigma(T_{g*}), \quad g = 1, 2, 3. \quad (9.12.3)$$

Equations (9.12.1) may now be written

$$\begin{aligned} y_{1a} &= A_1 \tau_{0a} + B_1 + e_{1a}, \\ y_{2a} &= A_2 \tau_{0a} + B_2 + e_{2a}, \\ y_{3a} &= A_3 \tau_{0a} + B_3 + e_{3a}. \end{aligned} \quad (9.12.4)$$

Consider the covariances between the observable variables. Since the errors are assumed to be uncorrelated with other variables, Eqs. (9.12.4) and (9.12.3) yield

$$\begin{aligned}\sigma(Y_{1*}, Y_{2*}) &= A_1 A_2 \sigma_T^2 = \sigma(T_{1*})\sigma(T_{2*}), \\ \sigma(Y_{1*}, Y_{3*}) &= A_1 A_3 \sigma_T^2 = \sigma(T_{1*})\sigma(T_{3*}), \\ \sigma(Y_{2*}, Y_{3*}) &= A_2 A_3 \sigma_T^2 = \sigma(T_{2*})\sigma(T_{3*}).\end{aligned}\quad (9.12.5)$$

We may solve these equations for  $\sigma(T_{1*})$ ,  $\sigma(T_{2*})$ , and  $\sigma(T_{3*})$ :

$$\begin{aligned}\sigma^2(T_{1*}) &= \frac{\sigma(Y_{1*}, Y_{2*})\sigma(Y_{1*}, Y_{3*})}{\sigma(Y_{2*}, Y_{3*})}, \\ \sigma^2(T_{2*}) &= \frac{\sigma(Y_{1*}, Y_{2*})\sigma(Y_{2*}, Y_{3*})}{\sigma(Y_{1*}, Y_{3*})}, \\ \sigma^2(T_{3*}) &= \frac{\sigma(Y_{1*}, Y_{3*})\sigma(Y_{2*}, Y_{3*})}{(\sigma Y_{1*}, Y_{2*})}.\end{aligned}\quad (9.12.6)$$

If we substitute the sample covariances of  $Y$ 's for the corresponding population parameters in (9.12.6), we obtain consistent estimates for the variances of the true scores. Then a consistent estimate  $\hat{\rho}_g$  of the reliability coefficient for test  $g$  is provided by

$$\hat{\rho}_g = \hat{\sigma}^2(T_{g*})/\hat{\sigma}^2(Y_{g*}), \quad g = 1, 2, 3, \quad (9.12.7)$$

where  $\hat{\sigma}(T_{g*})$  is the sample estimate of  $\sigma(T_{g*})$ .

We require three forms of the test to obtain the estimates so far discussed. If there are just three forms, the true-score variances are determined exactly by the available equations; thus the usual sample estimates for these parameters are appropriate. If there are more than three test forms, there are more equations relating the population parameters to each other than there are parameters; in this situation, special sample estimates with good properties may be obtained, as discussed below.

We see from (9.12.5) that when there are  $g = 1, 2, \dots, n$  test forms, column  $g$  of the variance-covariance matrix of the observed scores is proportional to any other column  $g'$ , provided that the elements in rows  $g$  and  $g'$  are ignored. This means that after appropriate modifications in the diagonal elements, the matrix will be of rank one. The problem of estimating  $\hat{\sigma}(T_{g*})$  under this model is thus a specialized problem in factor analysis (see Section 16.7). In particular, if the observed scores have a normal multivariate distribution, then Lawley's maximum likelihood procedures (Lawley and Maxwell, 1963; Jöreskog, in press) can be used to provide efficient estimates of  $\sigma(T_{g*})$ . They can also be used in statistical investigations designed to determine whether Eq. (9.12.2) provides credible approximations.

These assumptions may be relaxed further. Suppose that the true score on  $Y_g$  is not simply  $T_g$ , but rather  $T_g + T'_g$ , where  $T'_g$  is uncorrelated with  $T_g$ , and

with  $T'_{g'}$  for any  $g' \neq g$ . A repetition of the development already given shows that  $T'_g$  is indistinguishable from  $E_g$  for present purposes, and therefore the equations derived from (9.12.1) are valid for this broader case.

The foregoing assumptions may be relaxed still further. The relationship of  $T_2, T_3, \dots, T_n$  to  $T_1$  need not be assumed linear, but may be taken to be a polynomial of some specified degree  $K$ . If enough forms of the test are available, it will be possible to estimate not only the mean and variance of the true score of each different form but also higher true-score moments up to order  $2K$ , using methods similar to those described in Chapter 10.

## Exercises

- 9.1. Show that when  $\epsilon_{g*}$  and  $\zeta_*$  are positively correlated, then

$$\rho^2(Y_{g*}, \zeta_*) \geq 1 - \frac{\sigma^2(\epsilon_{g*})}{\sigma^2(Y_{g*})} \geq \frac{\sigma^2(\zeta_*)}{\sigma^2(Y_{g*})}.$$

Use the following procedure:

- a) Show that

$$1 - \frac{\sigma^2(\epsilon_{g*})}{\sigma^2(Y_{g*})} = \frac{\sigma^2(\zeta_*)}{\sigma^2(Y_{g*})} + 2 \frac{\sigma(\epsilon_{g*}, \zeta_*)}{\sigma^2(Y_{g*})}.$$

- b) Show that

$$\rho^2(Y_{g*}, \zeta_*) = 1 - \frac{\sigma^2(\epsilon_{g*})}{\sigma^2(Y_{g*})} [1 - \rho^2(\epsilon_{g*}, \zeta_*)]$$

by expressing (9.8.1) in terms of covariances, expanding the resulting expression, and substituting the expression in (a).

- c) Obtain the desired inequality by using (a) and (b).

- 9.2. Assume the data in Exercise 7.2 and calculate the following quantities:

- a) The variance components analysis estimates of the specific reliability coefficient. Use (1) all three replications, and (2) replications 1 and 2 only.  
 b) The correlation between replications 1 and 2. Explain why  $r^2(y_{g*1}, y_{g*2}) > \hat{\rho}_{YT}^2$  in this example while in Table 7.3.1 the inequality is reversed.

- 9.3. Assume the data in Exercise 7.2 and use formulas (9.2.1) and (9.2.2) to find an estimate for the specific reliability coefficient, given (1) all three replications, and (2) replications 1 and 2 only.

- 9.4. Assume the data of Exercise 7.2. Suppose there are no replications but three nominally parallel tests. Use (9.2.8) to find an estimate of a lower bound to the specific reliability coefficient for each test.

- 9.5. Assume the data of Exercise 7.2 and make the same assumption as in the previous problem. Find

- a) an estimate for the generic reliability coefficient, using (9.7.6) and the first two tests only, and  
 b) an estimate for the generic reliability coefficient for test 3, using all three tests.

### References and Selected Readings\*

- American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, *Standards for educational and psychological tests and manuals*. Washington, D.C.: American Psychological Association, 1966.
- ANDERSON, C. C., Function fluctuation. *British Journal of Psychology, Monograph Supplements*, 1958, **30**.
- ANDERSON, T. W., The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika*, 1963, **28**, 1-25.
- AOYAMA, H., Sampling fluctuations of the test reliability. *Annals of the Institute of Statistical Mathematics* (Tokyo), 1957, **8**, 129.
- BOCK, R. D., Contributions of multivariate experimental designs to educational research. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*, 1967, pp. 820-840.
- BOCK, R. D., Multivariate analysis of variance of repeated measurements. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963, pp. 85-103.
- BOCK, R. D., and E. A. HAGGARD, The use of multivariate analysis of variance in behavioral research. In *Handbook of measurement in educational psychology and education*. Reading, Mass.: Addison-Wesley, in press.
- BRILLINGER, D. R., The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. *Review of the International Statistical Institute*, 1964, **32**, 202-206.
- BUROS, O. K., *Schematization of old and new concepts of test reliability based upon parametric models*. New Brunswick, N.J.: Gryphon Press, 1963. (Dittoed).
- BURT, C., Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology*, 1955, **8**, 103-118.
- CHURCH, A., Jr., Analysis of data when the response is a curve. *Technometrics*, 1966, **8**, 229-246.
- COLE, J. W. L., and J. E. GRIZZLE, Applications of multivariate analysis of variance to repeated-measurements experiments. *Biometrics*, 1966, **22**, 810-828.
- COLEMAN, J. S., *Models of change and response uncertainty*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.
- COLLIER, R. O., JR., F. B. BAKER, and G. K. MANDEVILLE, Tests of hypothesis in a repeated measures design from a permutation viewpoint. *Psychometrika*, 1967, **32**, 15-24.
- CRAMÉR, H., *Mathematical methods of statistics*. Princeton, N.J.: Princeton University Press, 1946.
- CRONBACH, L. J., and GOLDINE C. GLESER, The signal noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, 1964, **24**, 467-480.

---

\* For basic references on reliability, also see those listed for Chapter 8.

- CRONBACH, L. J., H. IKEDA, and R. A. AVNER, Intraclass correlation as an approximation to the coefficient of generalizability. *Psychological Reports*, 1964, **15**, 727-736.
- CRONBACH, L. J., NAGESWARI RAJARATNAM, and GOLDINE C. GLESER, Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, **16**, 137-163.
- CURETON, E. E., Reliability and validity: basic assumptions and experimental designs. *Educational and Psychological Measurement*, 1965, **25**, 327-346.
- DANFORD, M. B., H. M. HUGHES, and R. C. McNEE, On the analysis of repeated-measurements experiments. *Biometrics*, 1960, **16**, 547-565.
- FLEISS, J. L., Assessing the accuracy of multivariate observations. *Journal of the American Statistical Association*, 1966, **61**, 403-412.
- FREEMAN, H., *Introduction to statistical inference*. Reading, Mass.: Addison-Wesley, 1963.
- GAITO, J., and D. E. WILEY, Univariate analysis of variance procedures in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963, pp. 60-84.
- GARSIDE, R. F., The measurement of function fluctuation. *Psychometrika*, 1958, **23**, 75-83.
- GEISSE, S., A method for testing treatment effects in the presence of learning. *Biometrics*, 1959, **15**, 389-395.
- GLASER, R., The reliability of inconsistency. *Educational and Psychological Measurement*, 1952, **12**, 60-64.
- GLESER, GOLDINE C., L. J. CRONBACH, and NAGESWARI RAJARATNAM, Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 1965, **30**, 395-418.
- GOLDFARB, N., *An introduction to longitudinal statistical analysis*. Glencoe, Ill.: Free Press, 1960.
- HOFFMAN, P. J., Test reliability and practice effects. *Psychometrika*, 1963, **28**, 273-288.
- HOLTZMAN, W. H., Statistical models for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963, pp. 199-211.
- HORN, J. L., and K. B. LITTLE, Isolating change and invariance in patterns of behavior. *Multivariate Behavioral Research*, 1966, **1**, 219-228.
- HOWARD, K. I., Differentiation of individuals as a function of repeated testing. *Educational and Psychological Measurement*, 1964, **24**, 875-894.
- JACKSON, R. W. B., and G. A. FERGUSON, *Studies on the reliability of tests*. Toronto: Department of Educational Research, University of Toronto. Bulletin No. 12, 1941.
- JÖRESKOG, K. G., Some contributions to maximum likelihood factor analysis. *Psychometrika*, in press.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. New York: Hafner, 1958-61, 2 vols.

- KRISTOF, W., Statistical inferences about the error variance. *Psychometrika*, 1963, **28**, 129-143. (a)
- KRISTOF, W., The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, 1963, **28**, 221-238. (b)
- KRISTOF, W., Testing differences between reliability coefficients. *British Journal of Statistical Psychology*, 1964, **17**, 105-111.
- LAFORGE, R., Components of reliability. *Psychometrika*, 1965, **30**, 187-195.
- LAWLEY, D. N., and A. E. MAXWELL, *Factor analysis as a statistical method*. London: Butterworth, 1963.
- LINDQUIST, E. F., Goodness of fit of trend curves and significance of trend differences. *Psychometrika*, 1947, **12**, 65-78.
- LINDQUIST, E. F., *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- LORD, F. M., A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika*, 1957, **22**, 207-220.
- McHUGH, R. B., and F. J. WALL, Estimating the precision of time period effects in longitudinal models with serially correlated and heterogeneous errors. *Biometrics*, 1962, **18**, 520-528.
- MADANSKY, A., Instrumental variables in factor analysis. *Psychometrika*, 1964, **29**, 105-113.
- MEDLEY, D. M., and H. E. MITZEL, Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand-McNally, 1963, pp. 247-328.
- MILLER, R. G., JR., A trustworthy jackknife. *Annals of Mathematical Statistics*, 1964, **35**, 1594-1605.
- NORMAN, W. T., On estimating psychological relationships: social desirability and self-report. *Psychological Bulletin*, 1967, **67**, 273-293.
- OLKIN, I., and J. W. PRATT, Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 1958, **29**, 201-211.
- POTTHOFF, R. F., and S. N. ROY, A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 1964, **51**, 313-326.
- RAJARATNAM, NAGESWARI, Reliability formulas for independent decision data when reliability data are matched. *Psychometrika*, 1960, **25**, 261-271.
- RAJARATNAM, NAGESWARI, L. J. CRONBACH, and GOLDINE C. GLESER, Generalizability of stratified-parallel tests. *Psychometrika*, 1965, **30**, 39-56.
- RAO, C. R., Statistical methods for comparison of growth curves. *Biometrics*, 1958, **14**, 1-17.
- RAO, C. R., The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 1965, **52**, 447-458.
- ROZEBOOM, W. W., *Foundations of the theory of prediction*. Homewood, Illinois: Dorsey Press, 1966.
- SCHAIE, K. W., A general model for the study of developmental problems. *Psychological Bulletin*, 1965, **64**, 92-107.

- STANLEY, J. C., Statistical analysis of scores from counterbalanced tests. *Journal of Experimental Education*, 1955, **23**, 187-208.
- STANLEY, J. C., Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika*, 1961, **26**, 205-219.
- STANLEY, J. C., Analysis-of-variance principles applied to the grading of essay tests. *Journal of Experimental Education*, 1962, **30**, 279-283.
- THORNDIKE, R. L., Intellectual status and intellectual growth. *Journal of Educational Psychology*, 1966, **57**, 121-127.
- TUCKER, L. R., Some mathematical notes on three-mode factor analysis. *Psychometrika*, 1966, **31**, 279-311.
- VINSONHALER, J. F., and W. MEREDITH, A stochastic model for repeated testing. *Multivariate Behavioral Research*, 1966, **1**, 461-478.
- WIGGINS, R. A., On factoring the correlation of discrete multivariable stochastic processes. Cambridge: Massachusetts Institute of Technology. Scientific Report No. 9, 1965.

# SOME TEST THEORY FOR $\tau$ -EQUIVALENT MEASUREMENTS, INCLUDING ESTIMATION OF HIGHER-ORDER MOMENTS\*

## 10.1 Introduction and Definitions

In classical test theory, the concept of parallel measurements is basic. In Chapter 8, on the other hand, we outlined a theory of nominally parallel tests—tests that are not parallel at all under the classical definition; in Chapter 11 we shall outline a mental test theory of “randomly parallel” tests. The present chapter outlines a theory for measurements that are not parallel but are essentially  $\tau$ -equivalent (see Definition 2.13.8). The observed scores  $Y_{1a}$ ,  $Y_{2a}, \dots, Y_{ga}, \dots$  are assumed to be essentially  $\tau$ -equivalent throughout this chapter.

We have deferred discussing the theory of essentially  $\tau$ -equivalent tests until now because it deals with moments of all orders. In the preceding chapters, we have been concerned with means, variances, and covariances, which are first- and second-order moments; here and in Chapter 11 we shall be concerned with higher-order moments as well. The main task of this chapter is to obtain unbiased estimates of the higher moments of the true-score and error distributions, and of the bivariate distribution of observed score and true score. Knowledge of these moments for any given set of data provides valuable information about the general nature of these distributions.

Here, as in Chapters 2 and 8, we assume that the examiner chooses to define the true score of examinee  $a$  as the examinee's expected score in an infinitely large population of measurements. If the observed score of examinee  $a$  on test  $g$ ,  $g = 1, 2, \dots$ , is denoted by  $Y_{ga}$ , then the true score of examinee  $a$  is here defined as

$$\xi_a \equiv \mathcal{E}_g Y_{ga}. \quad (10.1.1)$$

The error of measurement on test  $g$  (with respect to  $\xi_a$ ) is, by definition,

$$\epsilon_{ga} \equiv Y_{ga} - \xi_a. \quad (10.1.2)$$

Although  $\epsilon$  and  $\xi$  here have properties compatible with  $\epsilon$  and  $\xi$  in Chapters 8 and 9, we should not assume that results obtained in earlier chapters apply

---

\* This chapter deals with a specialized topic. It may be omitted without loss of continuity.

here. In this chapter, we start afresh with definitions (10.1.1) and (10.1.2). It need *not* be assumed that the errors for any fixed test  $g$  are unbiased. We shall develop a new model based on a weak assumption that we state in Section 10.2.

In the rest of this chapter, it will frequently be convenient to measure  $Y_{ga}$ ,  $\xi_a$ , and  $\epsilon_{ga}$  as deviations from their respective means  $\mathcal{E}_a Y_{ga}$ ,  $\mathcal{E}_a \xi_a$ , and  $\mathcal{E}_a \epsilon_{ga}$ . To keep the notation as readable as possible, we shall dispense with the use of upper- and lower-case letters to distinguish random variables from their numerical values, and use the lower-case letters  $y$ ,  $\tau$ , and  $e$  to denote variables or scores in deviation form, corresponding to the variables or scores  $Y$ ,  $\xi$ , and  $\epsilon$ . Thus, for each test  $g$  and examinee  $a$ ,

$$y_{ga} \equiv Y_{ga} - \mathcal{E}_a Y_{ga}, \quad \tau_a \equiv \xi_a - \mathcal{E}_a \xi_a, \quad e_{ga} \equiv \epsilon_{ga} - \mathcal{E}_a \epsilon_{ga}. \quad (10.1.3)$$

The substitution of  $\tau$  for  $\xi$  is intentional, since we shall find that  $\tau$  and  $e$  have much the same properties as the variables denoted by  $T$  and  $E$  in earlier chapters.

Note that the deviation score  $\tau_a$  so defined is the true score for the deviation score  $y_{ga}$  since

$$\tau_a = \mathcal{E}_g y_{ga}, \quad (10.1.4)$$

as the reader may readily verify. Also the deviation score  $e_{ga}$  is the error of measurement in  $y_{ga}$  since

$$y_{ga} = \tau_a + e_{ga}. \quad (10.1.5)$$

By (10.1.3),  $\mathcal{E}_a e_{ga} = 0$ ; that is, the errors  $e_{ga}$  are unbiased errors.

## 10.2 An Assumption of Linear Experimental Independence

For convenience, we might take the assumption underlying the present model to be that the conditional expectation over examinees of the errors in test  $g$  is always zero:

$$\mathcal{E}_a(e_{ga} | \tau_a; e_{ha}, e_{ia}, \dots) = 0, \quad (10.2.1)$$

for any tests  $g, h, i, \dots$ . Note that the distribution of  $e_{ga}$  is *not* to be assumed to be independent of  $\tau_a, e_{ha}, e_{ia}, \dots$ .

Actually Eq. (10.2.1) represents a somewhat stronger assumption than is required. However, (10.2.1) has more intuitive meaning than the minimum necessary assumption, which is

**Assumption 10.2.1.** *For some fixed value  $v$ ,*

$$\mathcal{E}_a(\tau_a^\lambda e_{fa} e_{ga}^\alpha e_{ha}^\beta \cdots e_{ja}^\delta) = 0, \quad (10.2.2)$$

where  $f, g, h, \dots, j$  represent any set of  $v$  tests and  $\lambda, \alpha, \beta, \dots, \delta$  are any  $v$  nonnegative integral exponents satisfying the restriction

$$\lambda + 1 + \alpha + \beta + \cdots + \delta \leq v. \quad (10.2.3)$$

We prove that Assumption 10.2.1 is implied by Eq. (10.2.1) as follows: If Eq. (10.2.1) holds, then for any  $v$  tests  $f, g, h, \dots, j$  and for any  $\lambda, \alpha, \beta, \dots, \delta$ ,

$$\mathcal{E}_a(e_{fa} | \tau_a; e_{ga}, e_{ha}, \dots, e_{ja}) = \mathcal{E}_a(e_{fa} | \tau_a^\lambda; e_{ga}^\alpha, e_{ha}^\beta, \dots, e_{ja}^\delta) = 0.$$

This last equation guarantees that  $e_{f*}$  is uncorrelated with the product

$$\tau_*^\lambda e_{g*}^\alpha e_{h*}^\beta \cdots e_{j*}^\delta;$$

that is, that

$$\mathcal{E}_a(e_{fa} \tau_a^\lambda e_{ga}^\alpha e_{ha}^\beta \cdots e_{ja}^\delta) = \mathcal{E}_a e_{fa} \cdot \mathcal{E}_a(\tau_a^\lambda e_{ga}^\alpha e_{ha}^\beta \cdots e_{ja}^\delta).$$

Since  $\mathcal{E}_a e_{fa}$  is zero, by (10.1.3), this last equation leads immediately to Eq. (10.2.2) and Assumption 10.2.1.

### 10.3 Immediate Implications

Assumption 10.2.1 subsumes the following familiar formulas:

$$\mathcal{E}_a e_{ga} = 0, \quad g = 1, 2, \dots, \quad (10.3.0)$$

$$\mathcal{E}_a e_{ga} \tau_a = 0 \quad \text{or} \quad \rho(e_{g*}, \tau_*) = 0 \quad \text{or} \quad \rho(e_{g*}, \xi_*) = 0, \quad g = 1, 2, \dots, \quad (10.3.1)$$

$$\mathcal{E}_a e_{ga} e_{ha} = 0 \quad \text{or} \quad \rho(e_{g*}, e_{h*}) = 0 \quad \text{or} \quad \rho(e_{g*}, \epsilon_{h*}) = 0, \quad g \neq h. \quad (10.3.2)$$

The foregoing equations involving  $e$  are effectively the same as three of the equations of Theorem 2.7.1, which gives the basic assumptions of classical test theory.

Just as in Chapter 3, these formulas lead to

$$\sigma^2(y_{g*}) = \sigma_\tau^2 + \sigma^2(e_{g*}), \quad (10.3.3)$$

$$\sigma(y_{g*}, y_{h*}) = \sigma_\tau^2. \quad (10.3.4)$$

*Formula (10.3.4) holds for essentially  $\tau$ -equivalent tests but does not hold for nominally parallel tests in general.*

A useful formula for the error variance for test  $g$  follows from (10.3.3) and (10.3.4):

$$\sigma^2(e_{g*}) = \sigma^2(y_{g*}) - \sigma(y_{g*}, y_{h*}). \quad (10.3.5)$$

*This formula expresses the error variance in terms of observed-score moments,* that is, in terms of quantities that can in principle be observed in a population of examinees and can in fact be estimated from sample data. The main objective of the present chapter is to obtain similar formulas expressing other central moments or cumulants\* of the unobservable variables  $\tau, e_g, e_h, \dots$  in terms of moments of the observed scores.

---

\* Cumulants are polynomial functions of moments having certain convenient invariance properties. Simple formulas are available for converting from cumulants to moments or from moments to cumulants; for example, see Kendall and Stuart (1958).

The error variance of test  $h$  depends on  $\sigma(y_{h*})$ , not on  $\sigma(y_{g*})$ :

$$\sigma^2(e_{h*}) = \sigma^2(y_{h*}) - \sigma(y_{g*}, y_{h*}). \quad (10.3.6)$$

*The present chapter does not assume that test forms  $g, h, \dots$  have the same error variances or the same observed-score variances:*

The fact of  $\tau$ -equivalence implies certain restrictions. Although  $\sigma(e_{g*})$  need not equal  $\sigma(e_{h*})$ , it is clear from (10.3.4) that if tests  $g, h, i, \dots$  are  $\tau$ -equivalent, then it must follow (see Exercise 3.7) that

$$\sigma(y_{g*}, y_{h*}) = \sigma(y_{g*}, y_{i*}) = \sigma(y_{h*}, y_{i*}) = \dots \quad (10.3.7)$$

Thus the assumption of  $\tau$ -equivalence can be partially checked for any given set of data by computing unbiased estimators of the covariances in (10.3.7). Further empirically verifiable implications of  $\tau$ -equivalence will be given by Eqs. (10.5.10), (10.5.11), (10.5.12), and (10.6.3).

#### 10.4 The Basic Theorem for $\tau$ -Equivalent Measurements

Since it will be necessary to deal with multivariate moments, the usual compact notation will (with minor modification) be used here. The symbol  $\mu_{\alpha\beta\dots\delta}$  will denote a  $\nu$ -variate central moment of the observed scores:

$$\mu_{\alpha\beta\dots\delta} = \mathcal{C}_a y_1^\alpha y_2^\beta \cdots y_\nu^\delta. \quad (10.4.1)$$

Similarly the symbol  $M_{\lambda,\alpha\beta\dots\delta}$  will denote a multivariate central moment of the true score and the errors of measurement on  $\nu$  tests:

$$M_{\lambda,\alpha\beta\dots\delta} = \mathcal{C}_a \tau_a^\lambda e_1^\alpha e_2^\beta \cdots e_\nu^\delta. \quad (10.4.2)$$

In this notation, the eight equations in Section 10.3 can be written, respectively,

$$M_{0,1} = 0 \quad \text{or} \quad M_{0,10} = 0, \quad (10.4.3)$$

$$M_{1,1} = 0 \quad \text{or} \quad M_{1,10} = 0, \quad (10.4.4)$$

$$M_{0,11} = 0, \quad (10.4.5)$$

$$\mu_{20} = M_{2,00} + M_{0,20}, \quad (10.4.6)$$

$$\mu_{11} = M_{2,00}, \quad (10.4.7)$$

$$M_{0,20} = \mu_{20} - \mu_{11}, \quad (10.4.8)$$

$$M_{0,02} = \mu_{02} - \mu_{11}, \quad (10.4.9)$$

$$\mu_{110} = \mu_{101} = \mu_{011}. \quad (10.4.10)$$

Equations (10.4.4), (10.4.5), (10.4.8), and (10.4.9) express the second-order moments of the latent variables; for the nonvanishing moments, the expressions are in terms of the second-order moments of  $\nu = 2$  observed-score variables.

The basic result of the present chapter is the following theorem, which promises similar formulas for higher-order moments and cumulants:

**Theorem 10.4.1.** *Under Assumption 10.2.1, each univariate and multivariate cumulant of the true score, defined by (10.1.1), and of error scores  $\epsilon_1, \epsilon_2, \dots, \epsilon_v$  up to order  $v$ , can be expressed as a linear function of cumulants of observed scores  $Y_1, Y_2, \dots, Y_v$ .*

The next two sections are concerned with the proof of this theorem.

### 10.5 Third-Order Moments

From (10.1.5),

$$y_{ga}^3 = \tau_a^3 + 3\tau_a^2 e_{ga} + 3\tau_a e_{ga}^2 + e_{ga}^3. \quad (10.5.1)$$

Taking the expectation over people, we obtain

$$\mu_3 = M_{3,0} + 3M_{2,1} + 3M_{1,2} + M_{0,3}.$$

By Assumption 10.2.1, the second term on the right vanishes. Adding some zero subscripts that do not change the meaning of the notation, we have

$$\mu_{300} = M_{3,000} + 3M_{1,200} + M_{0,300}. \quad (10.5.2)$$

Similarly

$$\begin{aligned} y_{ga}^2 y_{ha} &= (\tau_a + e_{ga})^2 (\tau_a + e_{ha}) \\ &= \tau_a^3 + 2\tau_a^2 e_{ga} + \tau_a e_{ga}^2 + \tau_a^2 e_{ha} + 2\tau_a e_{ga} e_{ha} + e_{ga}^2 e_{ha}. \end{aligned} \quad (10.5.3)$$

If we take the expectation over people, then

$$\mu_{21} = M_{3,00} + 2M_{2,10} + M_{1,20} + M_{2,01} + 2M_{1,11} + M_{0,21}.$$

All terms on the right are zero except the first and the third, so finally

$$\mu_{210} = M_{3,000} + M_{1,200}. \quad (10.5.4)$$

Similarly we find that

$$\mu_{111} = M_{3,000}. \quad (10.5.5)$$

Equations (10.5.2), (10.5.4), and (10.5.5) express each third-order observed-score central moment as a linear function of third-order true- and error-score moments. These equations can be solved by successive substitutions to obtain equations that express each nonvanishing true- and error-score central moment as a linear function of the observed-score moments:

$$M_{3,000} = \mu_{111}, \quad (10.5.6)$$

$$M_{1,200} = \mu_{210} - \mu_{111}, \quad (10.5.7)$$

$$M_{0,300} = \mu_{300} - 3\mu_{210} + 2\mu_{111}. \quad (10.5.8)$$

All other third-order central moments vanish, by Assumption 10.2.1.

By symmetry, there is a companion equation to (10.5.7) that must also be true. The two symmetrical equations are

$$M_{1,200} = \mu_{210} - \mu_{111}, \quad M_{1,200} = \mu_{201} - \mu_{111}. \quad (10.5.9)$$

From (10.5.9), it follows that

$$\mu_{210} = \mu_{201}. \quad (10.5.10)$$

This equation states the equality of two observed-score moments. By symmetry, again, it follows from (10.5.10) that

$$\mu_{120} = \mu_{021}, \quad \mu_{102} = \mu_{012}. \quad (10.5.11)$$

Equations (10.5.10) and (10.5.11) thus represent observable implications of Assumption 10.2.1.

It is clear from (10.5.5) that if more than three  $\tau$ -equivalent tests are available, then all moments like  $\mu_{111}$  must be equal. For four tests, this requirement would be written

$$\mu_{1110} = \mu_{1101} = \mu_{1011} = \mu_{0111}, \quad (10.5.12)$$

an equation similar to (10.3.7). If, in a set of actual data, unbiased estimators of the moments fail to satisfy equations like (10.5.10), (10.5.11), and (10.5.12) except for sampling fluctuations, then we must reject Assumption 10.2.1 for these data.

The parameters  $M_{3,000}$  and  $M_{0,300}$  are, respectively, the third moment of the true score and the third moment of an error of measurement. From these we can compute measures of the skewness of the true scores and of the errors of measurement,  $M_{3,000}/M_{2,000}^{3/2}$  and  $M_{0,300}/M_{0,200}^{3/2}$ , respectively.

The parameter  $M_{1,200}$  may be usefully reinterpreted by the following equations, in which  $\mathcal{E}_\tau$  denotes an expectation over all possible values of  $\tau_a$ , and  $\sigma_{e|\tau}^2$  denotes the conditional variance of  $e_g$  for  $\tau_* = \tau$ :

$$\begin{aligned} M_{1,200} &= \mathcal{E}_a \tau_a e_{ga}^2 = \mathcal{E}_\tau [\mathcal{E}_a (\tau_a e_{ga}^2 | \tau_a = \tau)] \\ &= \mathcal{E}_\tau [\tau \mathcal{E}_a (e_{ga}^2 | \tau_a = \tau)] = \mathcal{E}_\tau (\tau \sigma_{e|\tau}^2) = \text{Cov}_\tau (\tau, \sigma_{e|\tau}^2). \end{aligned} \quad (10.5.13)$$

The last equality sign holds because

$$\mathcal{E}_\tau \tau_* = \mathcal{E}_a \tau_a = 0.$$

The quantity

$$\text{Cov}_\tau (\tau, \sigma_{e|\tau}^2) = \text{Cov}_\xi (\xi, \sigma_{E|\xi}^2)$$

is the covariance between true score and conditional error variance. *If there is an overall tendency for the error variance of the test to increase or to decrease with the true-score level of the examinee, this tendency will produce a nonzero  $M_{1,200}$ . (If the error variance has a curvilinear relation to true score, this may or may not produce a nonzero  $M_{1,200}$ .)*

## 10.6 Higher-Order Moments and Cumulants

The second- and third-order cumulants are identical with the second- and third-order moments already studied. Instead of working with higher-order moments, we find it more convenient to work with higher-order cumulants, particularly in practical work where unbiased estimation seems desirable.

The relation of the fourth-order cumulants, denoted by  $\kappa$ , to the moments  $\mu$  is given by

$$\begin{aligned}\kappa_{40} &= \mu_{40} - 3\mu_{20}^2, \\ \kappa_{31} &= \mu_{31} - 3\mu_{20}\mu_{11}, \quad \kappa_{22} = \mu_{22} - \mu_{20}\mu_{02} - 2\mu_{11}^2, \\ \kappa_{211} &= \mu_{211} - \mu_{200}\mu_{011} - 2\mu_{110}\mu_{101}, \\ \kappa_{1111} &= \mu_{1111} - \mu_{1100}\mu_{0011} - \mu_{1010}\mu_{0101} - \mu_{1001}\mu_{0110}.\end{aligned}\quad (10.6.1)$$

These equations are general and hold for any variables.

The fourth-order true- and error-score cumulants (denoted by  $K$ ) can be expressed in terms of observed-score cumulants (denoted by  $\kappa$ ) by the same techniques used in the preceding section. The resulting equations are

$$\begin{aligned}K_{4,0000} &= \kappa_{1111} \\ K_{2,2000} &= \kappa_{2110} - \kappa_{1111}, \\ K_{1,3000} &= \kappa_{3100} - 3\kappa_{2110} + 2\kappa_{1111}, \\ K_{0,2200} &= \kappa_{2200} - \kappa_{2110} - \kappa_{1210} + \kappa_{1111}, \\ K_{0,4000} &= \kappa_{4000} - 4\kappa_{3100} + 6\kappa_{2110} - 3\kappa_{1111}.\end{aligned}\quad (10.6.2)$$

All other fourth-order cumulants vanish as a result of Assumption 10.2.1.

If tests are  $\tau$ -equivalent, then equalities such as the following must hold among the fourth-order observed-score cumulants:

$$\begin{aligned}\kappa_{310} &= \kappa_{301}, \quad \kappa_{130} = \kappa_{031}, \quad \text{etc.,} \\ \kappa_{2110} &= \kappa_{2101} = \kappa_{2011}, \quad \text{etc.,} \\ \kappa_{11110} &= \kappa_{11101} = \kappa_{11011} = \kappa_{10111} = \kappa_{01111}, \quad \text{etc.}\end{aligned}\quad (10.6.3)$$

The rule is that two observed-score cumulants must be the same if the first can be converted to the second by a permutation that shifts only the zero and the unit subscripts of the first.

The proofs of these results and of Theorem 10.4.1 for higher-order cumulants are given in Lord (1959).

## 10.7 Regression of True Score on Observed Score

Since  $y_{ga} = \tau_a + e_{ga}$ , then by (10.2.2), the regression of observed score on true score is

$$\mathcal{E}_a(y_{ga} | \tau_a) = \tau_a + \mathcal{E}_a(E_{ga} | \tau_a) = \tau_a, \quad (10.7.0)$$

which is a straight line with a slope (regression coefficient) of 1. On the other hand, there is no reason in test theory why the regression of true score on observed score should be a straight line. Even if this regression happens to be rectilinear for some particular group of examinees with a particular frequency distribution of true scores, it is unlikely that it is strictly rectilinear in a subgroup having a different distribution of true scores. The assumption of linear regression of true score on observed score is probably often a good approximation for practical purposes, but it is undesirable as a general assumption since it is likely to hold, if at all, only for special groups of examinees (see Sections 22.8 and 22.9).

If the regression of true score on observed score is (or can be approximated by) a polynomial, then the regression can in principle be determined (approximated) by the methods of this chapter. Suppose, for illustrative purposes, that this regression is a third-degree polynomial, i.e., that for some given test  $g$ ,

$$\tau_a = A + By_{ga} + Cy_{ga}^2 + Dy_{ga}^3 + \tilde{e}_{ga}, \quad (10.7.1)$$

where  $A, B, C, D$  are unknown parameters to be determined and  $\tilde{e}_{ga}$  is an unbiased error of estimation (*not* an error of measurement). It is well known in general regression theory that the error of estimation  $\tilde{e}_{ga}$  has zero mean and is uncorrelated with  $y_{g*}, y_{g*}^2, y_{g*}^3$ .

Taking the expectation over examinees of both sides of (10.7.1), we have

$$0 = A + C\mathcal{E}_a y_{ga}^2 + D\mathcal{E}_a y_{ga}^3. \quad (10.7.2)$$

Multiplying (10.7.1) by  $y_{ga}$  and then taking the expectation gives

$$\mathcal{E}_a y_{ga} \tau_a = B\mathcal{E}_a y_{ga}^2 + C\mathcal{E}_a y_{ga}^3 + D\mathcal{E}_a y_{ga}^4. \quad (10.7.3)$$

Similarly, multiplying (10.7.1) by  $y_{ga}^2$  or by  $y_{ga}^3$  gives, respectively,

$$\begin{aligned} \mathcal{E}_a y_{ga}^2 \tau_a &= A\mathcal{E}_a y_{ga}^2 + B\mathcal{E}_a y_{ga}^3 + C\mathcal{E}_a y_{ga}^4 + D\mathcal{E}_a y_{ga}^5, \\ \mathcal{E}_a y_{ga}^3 \tau_a &= A\mathcal{E}_a y_{ga}^3 + B\mathcal{E}_a y_{ga}^4 + C\mathcal{E}_a y_{ga}^5 + D\mathcal{E}_a y_{ga}^6. \end{aligned} \quad (10.7.4)$$

If the various moments in (10.7.2), (10.7.3), and (10.7.4) are known, then these equations constitute four linear equations that are readily solved for the four unknown parameters  $A, B, C$ , and  $D$ . Although the moments on the right are usually unknown, they can be estimated directly from a frequency distribution of observed scores. The moments on the left can then be rewritten as follows:

$$\mathcal{E}_a y_{ga} \tau_a = \mathcal{E}_a (\tau_a + e_{ga}) \tau_a = \mathcal{E}_a \tau_a^2 + \mathcal{E}_a e_{ga} \tau_a = M_{2,0}, \quad (10.7.5)$$

$$\mathcal{E}_a y_{ga}^2 \tau_a = \mathcal{E}_a \tau_a^3 + 2\mathcal{E}_a \tau_a^2 e_{ga} + \mathcal{E}_a \tau_a e_{ga}^2 = M_{3,0} + M_{1,2}, \quad (10.7.6)$$

$$\begin{aligned} \mathcal{E}_a y_{ga}^3 \tau_a &= M_{4,0} + 3M_{2,2} + M_{1,3} \\ &= K_{4,0} + 3K_{2,0}^2 + 3K_{2,2} + 3K_{2,0}K_{0,2} + K_{1,3}. \end{aligned} \quad (10.7.7)$$

Formulas (10.7.5) through (10.7.7), together with the results of Sections 10.5 and 10.6, make it possible to estimate the left-hand terms of (10.7.3) and (10.7.4) from actual test data and thus to estimate the (curvilinear) regression of true score on observed score for a particular test. Such a regression could be used to estimate the true score of any individual examinee from his observed score.

Similar methods can be used to estimate the regression curves relating the true scores for two tests that measure different psychological dimensions.

### 10.8 Implications, Applications, and Limitations

If some random variable  $X$  can take on only  $v + 1$  discrete values  $x_0, x_1, \dots, x_u, \dots, x_v$ , then the frequency distribution of  $x$  is completely determined by its first  $v$  moments  $\mu'_1, \mu'_2, \dots, \mu'_r, \dots, \mu'_v$  about the origin (David and Barton, 1962, p. 223; Riordan, 1958, p. 32). If the variables are continuous, then the situation is more complicated. Kendall and Stuart (1958, Section 3.34) show that if two distributions have the same moments up to order  $v$ , then they have the same best-fitting (least-squares) polynomial and thus

will, in a sense, be approximations to one another. We shall encounter many cases where, although we cannot determine a distribution function explicitly, we may ascertain its moments at least up to some order; and hence we shall be able to approximate to the distribution by finding another distribution of known form which has the same lower moments. In practice, approximations of this kind often turn out to be remarkably good, even when only the first three or four moments are equated.

True-score moments estimated by the methods of this chapter can be used to infer the shape of the true-score distribution. Practical use of this approach, however, is limited by the fact that estimates of fourth-order moments have uncomfortably large sampling errors, even when based on 2000 cases; for moments above the fourth, the situation is much worse (as a rough guide, John Tukey has suggested the number of cases be increased tenfold for each increase in the order of the moment to be computed\*). There is also the problem of deciding which of many distributions having the same first four or five moments should be used to represent the true-score distribution. Pearson curves can be used if only four moments are involved. Charlier and Edgeworth series may be used, but are often unsatisfactory for skewed distributions.

In principle, the bivariate distribution of true score and error could be estimated from bivariate moments, such as  $M_{1,200}$ , that are found by the methods of this chapter. However, as Kendall and Stuart (1958, Section 6.41) have stated, "No satisfactory method has yet been found of setting up families of bivariate frequency surfaces . . .". Difficulties of this sort motivate the development of strong true-score theories such as the one outlined in Chapter 23.

---

\* Personal communication.

Such theories provide the necessary families of bivariate and multivariate distributions, but at the expense of making much stronger assumptions about the nature of the data than are required in the present chapter.

Lord (1960) applied the methods of this chapter to the scores (number of right answers) on a vocabulary test that had been administered to four different groups of examinees. He drew the following conclusions about the particular test and groups under study:

1. The distribution of the errors of measurement was not independent of true score.
2. In the high-ability group, error variance decreased on the average as true score increased; in the low-ability group, error variance increased on the average as true score increased.
3. Errors of measurement were not normally distributed.
4. In the high-ability group, the distribution of the errors of measurement was negatively skewed; in the low-ability group, this distribution was positively skewed.

These results would ordinarily be expected from the true-score theories outlined in Chapter 11 (see Section 11.9) and in Chapter 23. The conclusions seem reasonable, or can be plausibly rationalized, as consequences of a *floor effect* and a *ceiling effect*, that is, as consequences of the fact that the number of right answers on a test of  $n$  items cannot be less than zero nor more than  $n$ . Some similar conclusions were obtained empirically by Mollenkopf (1949).

### **References and Selected Readings**

- AITKEN, A. C., *Determinants and matrices*, 3rd ed. New York: Interscience, 1944.
- DAVID, F. N., and D. E. BARTON, *Combinatorial chance*. New York: Hafner, 1962.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. Vol. I: *Distribution theory*. New York: Hafner, 1958.
- LORD, F. M., The joint cumulants of true values and errors of measurement. *Annals of Mathematical Statistics*, 1959, **30**, 1000–1004.
- LORD, F. M., An empirical study of the normality and independence of errors of measurement in test scores. *Psychometrika*, 1960, **25**, 91–104.
- MOLLENKOPF, W. G., Variation of the standard error of measurement. *Psychometrika*, 1949, **14**, 189–229.
- RIORDAN, J., *An introduction to combinatorial analysis*. New York: Wiley, 1958.

# ITEM SAMPLING IN TEST THEORY AND IN RESEARCH DESIGN\*

## 11.1 Introduction

This chapter deals with the case where the test score  $x_a$  of examinee  $a$  is the sum of the item scores  $y_{ga}$ ,  $g = 1, 2, \dots, n$ , and where the  $n$  test items are considered as a random sample from a population of items. This *item-sampling model* makes no other assumptions about the nature of the test. In spite of this, the model yields many important results not obtainable from the classical model. Some of these, and the general method for deriving them, are the subject of the present chapter.

It may be helpful to have in mind an example of a concrete situation where item sampling might be appropriate. Suppose, as a first approximation, that spelling ability is defined as the proportion of words from some specified standard dictionary that the examinee is able to spell correctly. Spelling tests are constructed by drawing random samples of  $n$  words from the dictionary. The  $n$  words comprising one test are read to the examinee, who is required to write down the spelling. The examinee's score is the number of words that he spells correctly, or more conveniently, the proportion of the  $n$  words that he spells correctly. Clearly this proportion is an unbiased estimator of his "spelling ability" as defined above. As a matter of fact, in the situation defined, it is a sufficient statistic for estimating the examinee's spelling ability.

It is true that more efficient spelling tests can be built by stratified sampling from a dictionary. In fact, almost any mental test may properly be considered a *stratified random sample* of items (usually called a *stratified sample*). By definition, such a sample is simply a set of random samples, each drawn from an appropriate stratum. Consideration of random item-sampling is therefore basic to the more complicated theory of stratified item-sampling. Some stratified item-sampling theory has been contributed by Rajaratnam *et al.* (1965), by Cronbach, Schönemann, and McKie (1965), by McKie (1965), and by Shoemaker (1966). Although we shall carry the basic theory further than they, we shall restrict ourselves to simple random samples.

---

\* Reading of this chapter can be omitted without loss of continuity.

One reason for using item sampling in certain situations is practical utility. In scientific work, one frequently draws a random sample, not because this is the most representative kind of sample, but because its statistical properties are well known, whereas the statistical properties of some possibly more representative but subjectively chosen sample would not be known. Similarly, testing situations sometimes require that we select items at random (with or without previous stratification) because this is the only way in which we can secure a firm basis for statistical inferences.

A possible objection to the item-sampling model (for example, see Loevinger, 1965) is that one does not ordinarily build tests by drawing items at random from a pool. There is, however, a similar and equally strong objection to classical test theory: Classical theory requires test forms that are strictly parallel, and yet no one has ever produced two strictly parallel forms for any ordinary paper-and-pencil test. Classical test theory is to be considered a useful idealization of situations encountered with actual mental tests. The assumption of random sampling of items may be considered in the same way. Further, even if the items of a particular test have not actually been drawn at random, we can still make certain interesting projections: We can conceive an item population from which the items of the test might have been randomly drawn and then consider the score the examinee would be expected to achieve over this population. The abundant information available on such expected scores enhances their natural interest to the examiner.

It is sometimes asserted that unless a given set of items is actually obtained by random sampling from a population of items, then it is erroneous to imagine a population of items from which the set might have been drawn. The following quotation from Cornfield and Tukey (1956), written in a broader context than that of mental test theory, gives an opposite viewpoint.

In almost any practical situation where analytical statistics is applied, the inference from the observations to the real conclusion has two parts, only the first of which is statistical. A genetic experiment on *Drosophila* will usually involve flies of a certain race of a certain species. The statistically based conclusions cannot extend beyond this race, yet the geneticist will usually, and often wisely, extend the conclusion to (a) the whole species, (b) all *Drosophila*, or (c) a larger group of insects. This wider extension may be implicit or explicit, but it is almost always present. If we take the simile of the bridge crossing a river by way of an island, there is a statistical span from the near bank to the island, and a subject-matter span from the island to the far bank. Both are important.

By modifying the observation program and the corresponding analysis of the data, the island may be moved nearer to or farther from the distant bank, and the statistical span may be made stronger or weaker. In doing this it is easy to forget the second span, which usually can only be strengthened by improving the science or art on which it depends. Yet a balanced understanding of, and choice among, the statistical possibilities requires constant attention to the second span. It may often be worth while to move the island nearer to the distant bank, at the cost of weakening the statistical span—particularly when the subject-matter span is weak.

In an experiment where a population of  $C$  columns was specified, and a sample of  $c$  columns was randomly selected, it is clearly possible to make analyses where

- 1) the  $c$  columns are regarded as a sample of  $c$  out of  $C$ , or
- 2) the  $c$  columns are regarded as fixed.

The question about these analyses is not their validity but their wisdom. Both analyses will have the same mean, and will estimate the effects of rows identically. Both analyses will have the same mean squares, but will estimate the accuracy of their estimated effects differently. The analyses will differ in the length of their inferences; both will be equally strong statistically. Usually it will be best to make analysis (1) where the inference is more general. Only if this analysis is entirely unrevealing on one or more points of interest are we likely to be wise in making analysis (2), whose limited inferences may be somewhat revealing.

But what if it is unreasonable to regard  $c$  columns as any sort of a fair sample from a population of  $C$  columns with  $C > c$ . We can (at least formally and numerically) carry out an analysis with, say,  $C = \infty$ . What is the logical position of such an analysis? It would seem to be much as follows: We cannot point to a specific population from which the  $c$  columns were a random sample, yet the final conclusion is certainly not to just these  $c$  columns. We are likely to be better off to move the island to the far side by introducing an unspecified population of columns "like those observed" and making the inference to the mean of this population. This will lengthen the statistical span at the price of leaving the location of the far end vague. Unless there is a known, fixed number of reasonably possible columns, this lengthening and blurring is likely to be worth while.\*

## 11.2 Matrix Sampling

When a sample of  $n$  test items has been administered to a sample of  $N$  examinees, the available data are the item responses  $y_{ga}$ ,  $g = 1, 2, \dots, n$ ,  $a = 1, 2, \dots, N$  (this  $y_{ga}$  is not a deviation score, as in Chapter 10). We assume that the sample of items and the sample of examinees are drawn independently of each other. This kind of sampling is called *matrix sampling*; it has been discussed by Wilks (1962, Section 8.6).

For given  $g$  and  $a$ , we can consider  $Y_{ga}$  a random variable over replications or over the propensity distribution, as in earlier chapters. When replications cannot be observed, however, we can simplify our thinking for many purposes without arriving at incorrect conclusions. In this case, we shall no longer consider the observed score of examinee  $a$  on test  $g$  a random variable but simply an observed constant  $y_{ga}$ . We shall think of  $Y$  as a random variable over both people and tests, taking on particular values denoted by  $y_{ga}$ .

We may summarize the item responses and related statistics and parameters in matrices as shown below. Note that a matrix or vector is represented by its

---

\* From Cornfield, J., and J. W. Tukey, Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, 27, 907–949. Used by permission.

typical element enclosed in  $\| \quad \|$  or  $\{ \quad \}$ , respectively.

$$\begin{aligned} \|y_{ga}\| &\equiv \begin{vmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nN} \end{vmatrix}, \quad \left\| \frac{1}{N} \sum_{a=1}^N y_{ga} \right\| \equiv \begin{vmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{vmatrix}, \quad \|\mathcal{E}_A Y_{gA}\| \equiv \begin{vmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{vmatrix}, \\ \left\| \frac{1}{n} \sum_{g=1}^n y_{ga} \right\| &\equiv \|z_1 \ z_2 \ \cdots \ z_N\|, \\ \|\mathcal{E}_G Y_{Ga}\| &\equiv \|\xi_1 \ \xi_2 \ \cdots \ \xi_N\|. \end{aligned}$$

Lower-case subscripts range over the sample:  $g = 1, 2, \dots, n$  or  $a = 1, 2, \dots, N$ . Upper-case subscripts range over the population:  $G = 1, 2, \dots, \bar{n}$  or  $A = 1, 2, \dots, \bar{N}$ , where  $\bar{n}$  is the number of items in the population of items and  $\bar{N}$  is the number of examinees in the population of examinees. In most applications,  $\bar{n}$  and  $\bar{N}$  will be taken as infinite; however, many of the formulas will be developed for the more general case where the populations of items and examinees may be finite.

The matrix  $\|y_{ga}\|$  represents the responses of a sample of examinees drawn at random from a population of examinees to a sample of items drawn independently and at random from a population of items. All samples are drawn without replacement. The responses of all examinees in the population of examinees to all the items in the population of items are summarized in a matrix denoted by  $\|y_{GA}\|$ . The matrix sample  $\|y_{ga}\|$  is a submatrix of  $\|y_{GA}\|$ ; it is the matrix that remains after all but  $n$  rows and  $N$  columns of  $\|y_{GA}\|$  are crossed out.

The column vector  $\mathbf{p} = \{\sum_a y_{ga}/N\}$  has typical element  $p_i$ , the sample item difficulty for this sample of examinees; the column vector  $\boldsymbol{\pi} = \{\mathcal{E}_A Y_{gA}\}$  has typical element  $\pi_g$ , the item difficulty in the population of examinees. (Note that the term *difficulty* can be used for nondichotomous items.) The row vector  $\mathbf{z}' = \{\sum_g y_{ga}/n\}'$  has typical element  $z_a$ , the proportion of items in the test answered correctly by examinee  $a$ ; this is frequently referred to as the *relative observed score* of examinee  $a$ . The row vector  $\boldsymbol{\xi}' = \{\mathcal{E}_G Y_{Ga}\}'$  has typical element  $\xi_a = \mathcal{E}_G y_{Ga} = \mathcal{E}_G z_a$ , the *relative true score* of examinee  $a$ .

The reader should note that throughout the present chapter the *true score*  $\xi$  is a *relative true score*, that is, the expected score over all items in the population. It is a random variable taking values  $\xi_a$ . Note that  $\xi$  is the generic true score for a single item; the generic true score for a composite test composed of a random sample of  $n$  items is equal to  $n\xi$ . On occasion, we shall use the notations  $\xi \equiv n\xi$  and  $x = nz$ .

**Estimation problems.** In many of the usual, simple types of estimation problem, a population is completely specified by a convenient univariate frequency distribution. Many powerful estimation methods are available for such problems. Similarly a matrix population could be specified by an  $\bar{n}$ -variate or an  $\bar{N}$ -variate

frequency distribution, provided that an appropriate and convenient mathematical form could be found. However, with the exception of the multinomial and multivariate normal distributions and of distributions involving  $\bar{n}$  or  $\bar{N}$  independent variables, few simple multivariate distributions are known and none seems adequate for describing the types of matrix populations that are of concern here.

In the absence of an adequate parametric form for a frequency distribution, how are we to describe a matrix population without using a huge number of parameters? Hooke (1956a, 1956b) used either of two alternative sets of quantities, both of which are related to moments; one set is known as *generalized symmetric means* (*gsm*'s), the other as *bipolykays*. We shall define these quantities below as needed. The bipolykays are themselves linear functions of the *gsm*'s (the converse is also true). We can express many other quantities of interest as linear functions of *gsm*'s, and thus of the bipolykays; for example, the central moments of the true-score distribution (see Section 11.5). Conversely we can express the *gsm*'s and bipolykays as polynomial functions of appropriate moments and momentlike quantities (see Section 11.4).

Just as there are population *gsm*'s and bipolykays that characterize a matrix population, so there are sample *gsm*'s and bipolykays that characterize a matrix sample. One reason for choosing the *gsm*'s (and bipolykays) from among many other possible sets of descriptive quantities is that any given sample *gsm* (bipolykay) is an unbiased estimate of the same population *gsm* (bipolykay). Thus *an unbiased sample estimate for any population parameter is automatically available if the parameter can be expressed as a linear function of gsm's (or as a polynomial function of moments)*. Hooke (1956a, 1956b) gives most of the basic formulas.

### 11.3 Generalized Symmetric Means

We shall denote items by subscripts  $g, h, i, j$  or  $G, H, I, J$ , and examinees by  $a, b, c, d$  or  $A, B, C, D$ . Consider the expression

$$\frac{1}{M} \sum^M \neq y_{ga}^\alpha y_{gb}^\beta \cdots y_{gd}^\delta y_{ha}^\kappa y_{hb}^\lambda \cdots y_{jd}^\omega. \quad (11.3.1)$$

The Greek letters represent any set of nonnegative integral exponents (here, usually, these exponents will be 1). The symbol  $(\neq)$  on the summation sign indicates summation over all subscripts, except that row (item) subscripts represented by different letters must remain unequal throughout the summation, and similarly for column (examinee) subscripts;  $M$  indicates the total number of terms under the summation sign. We are here considering a finite sample of items and a finite sample of people; if either were infinite, expectations rather than summations would be used.

The expression (11.3.1) is called a *generalized symmetric mean* (*gsm*). This name arises from the fact that the expression is symmetric with respect to

column subscripts and also with respect to row subscripts. For example, if the first examinee (or test item) is replaced by the third and the third examinee (or test item) is replaced by the first, the value of a gsm is not altered. We may see this from the following simplified example, in which  $n = 4$  and the second subscript can take only the single value zero:

$$\frac{1}{12} \sum_{g,h=1}^{12} y_{go} y_{ho} = \frac{1}{12} (y_{10} y_{20} + y_{20} y_{10} + y_{10} y_{30} + y_{30} y_{10} + y_{10} y_{40} + y_{40} y_{10} \\ + y_{20} y_{30} + y_{30} y_{20} + y_{20} y_{40} + y_{40} y_{20} + y_{30} y_{40} + y_{40} y_{30}).$$

The gsm is a scalar quantity, not a matrix. A rather complicated example of a gsm is

$$\frac{1}{M} \sum_{g,h,i=1}^M y_{ga}^{\alpha} y_{gb}^{\beta} y_{ha}^{\gamma} y_{hc}^{\delta} y_{id}^{\epsilon} = \frac{1}{n^{[3]} N^{[4]}} \sum_{(g,h,i \neq)}^n \sum_{(a,b,c,d \neq)}^N y_{ga}^{\alpha} y_{gb}^{\beta} y_{ha}^{\gamma} y_{hc}^{\delta} y_{id}^{\epsilon}, \quad (11.3.2)$$

where  $n^{[r]} \equiv n(n - 1) \cdots (n - r + 1)$ . The *degree* of a gsm is the sum of the exponents of the  $y$ ; in (11.3.2), the degree is  $\alpha + \beta + \gamma + \delta + \epsilon$ .

The nature of a gsm is best understood, not by writing the entire sum out term by term, but rather by expressing the gsm's in terms of more familiar symmetric functions. We give one simple example:

$$\begin{aligned} \sum_{g,h=1}^{[2]} y_{ga} y_{ha} &= \sum_{g=1}^n \sum_{h=1}^n \sum_{a=1}^N y_{ga} y_{ha} - \sum_{g=1}^n \sum_{a=1}^N y_{ga}^2 \\ &= \sum_{a=1}^N \left( \sum_{g=1}^n y_{ga} \right) \left( \sum_{h=1}^n y_{ha} \right) - \sum_{g=1}^n \sum_{a=1}^N y_{ga}^2 \\ &= \sum_{a=1}^N \left( \sum_{g=1}^n y_{ga} \right)^2 - \sum_{g=1}^n \sum_{a=1}^N y_{ga}^2. \end{aligned} \quad (11.3.3)$$

Hooke uses a simplified notation for a generalized symmetric mean. He replaces the gsm by a pair of square brackets that enclose a matrix. This matrix has one row for each distinct row subscript ( $g, h$ , and  $i$  in Eq. 11.3.2) and one column for each distinct column subscript ( $a, b, c$ , and  $d$ ). Each entry in the matrix is the exponent on the corresponding  $y$ . Thus we may write the gsm in (11.3.2)

$$\begin{bmatrix} \alpha & \beta & & \\ \gamma & & \delta & \\ & & & \epsilon \end{bmatrix}.$$

The rows of this  $3 \times 4$  matrix have been taken to correspond to the subscripts  $g, h$ , and  $i$ , in that order; the columns, to  $a, b, c, d$ . Thus the  $\delta$  in the second row and third column indicates that the term  $y_{hc}$  in the gsm has an exponent of  $\delta$ . In this square-bracket symbolism, zero entries need not be written down. Permutation of columns (rows) is equivalent to exchanging column (row) subscripts; such manipulations do not change the meaning of the bracket.

We shall replace brackets having only one row and one column by larger brackets containing at least one null row and one null column. For example,

$$[1] \equiv \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \equiv \frac{1}{nN} \sum_{g=1}^n \sum_{a=1}^N y_{ga} = \frac{1}{N} \sum_a z_a = \bar{z} = \frac{\bar{x}}{n}$$

and

$$[2] \equiv \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \equiv \frac{1}{nN} \sum_g \sum_a y_{ga}^2.$$

Thus the superscript in  $n^{[2]}$  cannot be misread as a gsm.

The generalized symmetric means have one key property: *The expected value of the sample gsm over all possible matrix samples is simply the corresponding population gsm.* This property of the gsm's has been summarized briefly by saying that they are *inherited on the average*. We write the expected value of a gsm in matrix sampling from a finite matrix population simply by replacing lower-case subscripts by upper-case subscripts,  $n$  by  $\bar{n}$ , and  $N$  by  $\bar{N}$ . Thus, for example, denoting the expectation over all possible matrix samples by  $\mathcal{E}_G \mathcal{E}_A$  (or  $\mathcal{E}_I \mathcal{E}_A$ ), we have

$$\mathcal{E}_G \mathcal{E}_A \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{\bar{n}\bar{N}} \sum_{G=1}^{\bar{n}} \sum_{A=1}^{\bar{N}} y_{GA} = \frac{1}{\bar{N}} \sum_{A=1}^{\bar{N}} \xi_A = \mu_\xi \equiv \mu.$$

If either  $\bar{n}$  or  $\bar{N}$  is infinite, the symbol

$$\frac{1}{\bar{n}} \sum^{\bar{n}} \quad \text{or} \quad \frac{1}{\bar{N}} \sum^{\bar{N}}$$

must be replaced by an expectation.

Quantities like

$$\frac{1}{N} \sum_a y_a, \quad \frac{1}{nN} \sum_a \sum_i y_{ai}, \quad \frac{1}{N^{[2]}} \sum_{a \neq b} \sum y_{ay_b}, \quad \frac{1}{nN^{[2]}} \sum_{a,b} \neq \sum_i y_{ai} y_{bi}$$

are inherited on the average; quantities like

$$\left( \frac{1}{N} \sum_a y_a \right)^2 = \frac{1}{N^2} \sum_a \sum_b y_{ay_b} \quad \text{and} \quad \frac{1}{nN^2} \sum_a \sum_b \sum_i y_{ai} y_{bi}$$

are *not* inherited on the average. The reason is that the last two summations allow  $a = b$ ; thus they contain such terms as  $y_a^2$  and  $y_{ai}^2$ , which have different expectations than do  $y_{ay_b}$  and  $y_{ai} y_{bi}$ , respectively, when  $b \neq a$ . For example, in random sampling of examinees from a finite population, the expected value of the product  $y_{ay_b}$  over all possible pairs of examinees, allowing  $a = b$ , is simply

$$\frac{1}{\bar{N}^2} \sum_{A=1}^{\bar{N}} \sum_{B=1}^{\bar{N}} y_{AY_B} = \frac{1}{\bar{N}^2} \left( \sum_{A=1}^{\bar{N}} y_A \right) \left( \sum_{B=1}^{\bar{N}} y_B \right) = \frac{1}{\bar{N}^2} \left( \sum_{A=1}^{\bar{N}} y_A \right)^2.$$

This is not the same as the expectation over all possible pairs with  $a \neq b$ , which is

$$\begin{aligned} \frac{1}{N^{[2]}} \sum_{A \neq B}^N y_A y_B &= \frac{1}{\bar{N}^{[2]}} \left( \sum_{A=1}^{\bar{N}} \sum_{B=1}^{\bar{N}} y_A y_B - \sum_{A=1}^{\bar{N}} y_A^2 \right) \\ &= \frac{1}{\bar{N}^{[2]}} \left( \sum_{A=1}^{\bar{N}} y_A \right)^2 - \frac{1}{\bar{N}^2} \sum_{A=1}^{\bar{N}} y_A^2. \end{aligned}$$

#### 11.4 First- and Second-Degree gsm's

The one and only first-degree gsm, as we have already noted in the preceding section, is

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \equiv \frac{1}{nN} \sum_{g=1}^n \sum_{a=1}^N y_{ga} = \frac{1}{N} \sum_a z_a = \bar{z} = \frac{\bar{x}}{n}. \quad (11.4.1)$$

Furthermore, by inheritance on the average,

$$\mathcal{E}_G \mathcal{E}_A \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{\bar{n}\bar{N}} \sum_{G=1}^{\bar{n}} \sum_{A=1}^{\bar{N}} y_{GA} \equiv \mu_\xi, \quad \text{so that} \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

is an unbiased estimate of  $\mu_\xi$ , the population mean relative true score. In this connection, note that  $\mu_\pi \equiv \mathcal{E}_G \pi_G = \mathcal{E}_G \mathcal{E}_A Y_{GA} = \mu_\xi$ , and similarly that

$$\bar{p} \equiv \frac{1}{n} \sum_g p_g = \bar{z}.$$

The simplest second-degree gsm is

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \equiv \frac{1}{nN} \sum_g \sum_a y_{ga}^2. \quad (11.4.2)$$

If the test items are binary, i.e., if they can be scored only zero or one, it will be convenient to use the letter  $u$  instead of  $y$ , with the understanding that  $u_{ga}$  is always zero or one. For binary items, (11.4.2) becomes

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{nN} \sum_g \sum_a u_{ga}^2 = \frac{1}{nN} \sum_g \sum_a u_{ga} = \frac{1}{n} \sum_g p_g = \bar{z} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \quad (11.4.3)$$

Thus, in the binary case, the set of second-degree gsm's includes the first-degree gsm.

It is more laborious to rewrite the next gsm,

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \equiv \frac{1}{n^{[2]} N} \sum_{g \neq h}^n \sum_{a=1}^{n^{[2]}} y_{ga} y_{ha}, \quad (11.4.4)$$

in terms of familiar test-score and item statistics. We shall do this here only

for the binary case:

$$\begin{aligned}
 \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} &= \frac{1}{n^{[2]}N} \sum_{g \neq h}^n \sum_{a=1}^{[2]} u_{ga} u_{ha} \\
 &= \frac{1}{n^{[2]}N} \sum_{a=1}^N \left( \sum_{g=1}^n \sum_{h=1}^n u_{ga} u_{ha} - \sum_{g=1}^n u_{ga}^2 \right) \\
 &= \frac{1}{n^{[2]}N} \sum_a \left[ \left( \sum_g u_{ga} \right) \left( \sum_h u_{ha} \right) - \sum_g u_{ga} \right] \\
 &= \frac{1}{n^{[2]}} \left( \frac{1}{N} \sum_a x_a^2 - \frac{1}{N} \sum_a x_a \right) = \frac{1}{n^{[2]}} (s_x^2 + \bar{x}^2 - \bar{x}) \\
 &= \frac{1}{n-1} (ns_z^2 + n\bar{z}^2 - \bar{z}),
 \end{aligned} \tag{11.4.5}$$

where  $s^2$  denotes a sample variance and a superior bar denotes a sample mean.

By inheritance on the average, the expected value of this gsm for the case of binary items is

$$\mathcal{E}_G \mathcal{E}_A \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \frac{1}{\bar{n}^{[2]}N} \sum_{G \neq H}^{\bar{n}^{[2]}} \sum_{A=1}^{\bar{N}} u_{GA} u_{HA} = \frac{1}{\bar{n}-1} (\bar{n}\sigma_\xi^2 + \bar{n}\mu_\xi^2 - \mu_\xi), \tag{11.4.6}$$

where  $\mu_\xi$  and  $\sigma_\xi^2$  are the population mean and variance of the relative true score  $\xi$ . If  $\bar{n} \rightarrow \infty$ , so that the population of items becomes infinitely large, then by a limiting process,

$$\mathcal{E}_G \mathcal{E}_A \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \sigma_\xi^2 + \mu_\xi^2. \tag{11.4.7}$$

It may be shown that this equation holds when  $\bar{n} \rightarrow \infty$ , whether or not the items are binary.

From considerations of symmetry,

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \frac{1}{N-1} (Ns_p^2 + N\bar{z}^2 - \bar{z}) \tag{11.4.8}$$

for binary items, where  $s_p^2$  is the sample variance over items of the sample item difficulties. When the population of examinees is infinite,

$$\mathcal{E}_G \mathcal{E}_A \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \sigma_\pi^2 + \mu_\xi^2, \tag{11.4.9}$$

where  $\sigma_\pi^2$  is the population variance of the population item difficulties.

The only remaining second-degree gsm is

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \equiv \frac{1}{n^{[2]}N^{[2]}} \sum_{g \neq h}^n \sum_{a=1}^{[2]} \sum_{b=1}^{[2]} y_{ga} y_{hb}. \tag{11.4.10}$$

For binary items,

$$\begin{aligned}
 & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\
 &= \frac{1}{n^{[2]} N^{[2]}} \left[ \sum_{g=1}^n \sum_{h=1}^n \sum_{a=1}^N \sum_{b=1}^N u_{ga} u_{hb} - \sum_{g=1}^n \sum_{a=1}^N \sum_{b=1}^N u_{ga} u_{gb} \right. \\
 &\quad \left. - \sum_{g=1}^n \sum_{h=1}^n \sum_{a=1}^N u_{ga} u_{ha} + \sum_{g=1}^n \sum_{a=1}^N u_{ga}^2 \right] \\
 &= \frac{1}{n^{[2]} N^{[2]}} \left[ \left( \sum_g \sum_a u_{ga} \right) \left( \sum_h \sum_b u_{hb} \right) - N^2 \sum_g p_g^2 - n^2 \sum_a z_a^2 + \sum_g \sum_a u_{ga} \right] \\
 &= \frac{1}{n^{[2]} N^{[2]}} [n^2 N^2 \bar{z}^2 - n N^2 (s_p^2 + \bar{p}^2) - n^2 N (s_z^2 + \bar{z}^2) + N n \bar{z}] \\
 &= \frac{1}{(n-1)(N-1)} [(nN - N - n)\bar{z}^2 - N s_p^2 - n s_z^2 + \bar{z}]. \tag{11.4.11}
 \end{aligned}$$

The expected value can be written directly from the last expression in (11.4.11):

$$\begin{aligned}
 \mathcal{E}_I \mathcal{E}_A \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} &= \frac{1}{(\bar{n}-1)(\bar{N}-1)} [(\bar{n}\bar{N} - \bar{N} - \bar{n})\mu^2 - \bar{N}\sigma_\pi^2 - \bar{n}\sigma_\xi^2 + \mu] \\
 &= \mu^2 - \frac{1}{(\bar{n}-1)(\bar{N}-1)} [\bar{N}\sigma_\pi^2 + \bar{n}\sigma_\xi^2 - \mu(1-\mu)]. \tag{11.4.12}
 \end{aligned}$$

For  $\bar{n} \rightarrow \infty$  and  $\bar{N} \rightarrow \infty$ , as is frequently appropriate, it follows that

$$\mathcal{E}_I \mathcal{E}_A \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mu_\xi^2. \tag{11.4.13}$$

We can see from the derivation that this last equation holds whether or not the items are binary.

It is important to note that although  $\bar{x}$  is an unbiased estimate of  $\mu_x = \mu_x/n$ ,  $\bar{x}^2$  is *not* an unbiased estimate of  $\mu_x^2 = \mu_\xi^2/n^2$ . This is true even in ordinary (examinee) sampling. Whereas  $\mathcal{E}_A \bar{x} \equiv \mu_x$ , the expected value of  $\bar{x}^2$  in a sample of  $N$  cases is

$$\mathcal{E}_A \bar{x}^2 = \mathcal{E}_A \bar{x}^2 - (\mathcal{E}_A \bar{x})^2 + (\mathcal{E}_A \bar{x})^2 = \text{Var}_A(\bar{x}) + (\mathcal{E}_A \bar{x})^2 = \mu_x^2 + \sigma_x^2/N, \tag{11.4.14}$$

where  $\text{Var}_A(\bar{x}) = \sigma_x^2/N$  is the usual (examinee-) sampling variance of  $\bar{x}$ . Consequently

$$\mathcal{E}_A \bar{x}^2 \geq \mu_x^2, \tag{11.4.15}$$

with equality occurring only in the case when the  $x$  in the population are all equal. When  $\bar{n} \rightarrow \infty$  and  $\bar{N} \rightarrow \infty$ , an unbiased estimate of  $\mu_\xi^2$  in matrix sampling is given by the gsm of (11.4.11), as shown in (11.4.13).

Equations (11.4.1), (11.4.3), (11.4.5), (11.4.8), and (11.4.11) express the first- and second-degree gsm's in terms of certain central moments. These equations may be solved to express each central moment as a linear function of gsm's. The resulting formulas are:

$$\bar{z}^2 = \frac{1}{nN} \left\{ (n-1)(N-1) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (N-1) \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} + (n-1) \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\}, \quad (11.4.16)$$

$$\bar{s}_z^2 = \frac{N-1}{nN} \left\{ -(n-1) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (n-1) \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\}, \quad (11.4.17)$$

$$\bar{s}_p^2 = \frac{n-1}{nN} \left\{ -(N-1) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (N-1) \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\}. \quad (11.4.18)$$

With the aid of these equations, we can express any linear function of these central moments as a linear function of gsm's. We can also express products of these central moments as linear functions of gsm's, or of bipolykays, with the help of a multiplication table provided by Hooke (1956a, Table 3); but this is beyond the scope of this chapter. Once a function of parameters has been written as a linear function of population gsm's, an unbiased estimator of the function is automatically obtained by replacing the population gsm's by sample gsm's.

Most of the problems involving first- and second-degree moments considered in this chapter can be dealt with by using known formulas for analysis of variance components. The use of gsm's can simplify and systematize the solution of such problems. For problems involving third- and higher-degree moments, for example, for estimating the sampling variance of variance components without normality assumptions, the gsm or some other similar approach is virtually indispensable.

*Each central or noncentral moment of  $y$ ,  $x$ , or  $p$  of order  $r$  can be expressed as a linear function of gsm's of degree  $r$ .* The following table shows the number of different gsm's for degrees 1, 2, 3, and 4.

Degree	General case	Binary items
First	1	1
Second	4	3
Third	10	6
Fourth	33	16

Binary variables have fewer different gsm's because the exponents (denoted by  $\alpha, \beta, \dots, \omega$  in Eq. 11.3.1) do not have any effect when the variable is  $u_{ga} = 0, 1$ .

To express gsm's of degree  $r > 2$  in terms of more familiar quantities, we need certain momentlike quantities that involve  $r$ -tuple summations as well as the moments up through degree  $r$ . For binary items, for example, we can express all third-degree sample gsm's in terms of the first three moments of  $x$ , the first three moments of  $p$ , and the sums

$$\sum_g \sum_h p_g p_{gh} \quad \text{and} \quad \sum_g \sum_h \sum_i p_{ghi},$$

where  $p_{gh}$  is the proportion of examinees answering both items  $g$  and  $h$  correctly, and  $p_{ghi}$  is the proportion answering all three items  $g$ ,  $h$ , and  $i$  correctly. Some short-cut methods are available for computing such sums. Lord (1960, Table 1) gives computing formulas for all gsm's up through the fourth degree for binary items with  $N \rightarrow \infty$ . More general formulas can be worked out with the aid of David, Kendall, and Barton, *Symmetric Function and Allied Tables* (1966).

### 11.5 Estimating True-Score Moments

Consider the problem of finding an estimate for  $\sigma_\xi^2$  that is unbiased in matrix sampling. We see from (11.4.7) and (11.4.13) that when  $\bar{n} \rightarrow \infty$  and  $\bar{N} \rightarrow \infty$ , then

$$\mathcal{E}_I \mathcal{E}_A \left\{ \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} = \sigma_\xi^2.$$

Consequently

$$\hat{\sigma}_\xi^2 \equiv \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (11.5.1)$$

is an unbiased estimate of  $\sigma_\xi^2$ . This may be rewritten with the aid of (11.4.5) and (11.4.11) to show that *for binary items and for matrix sampling from infinite populations of items and examinees, an unbiased estimate of the variance of relative true score is given by*

$$\hat{\sigma}_\xi^2 = \frac{1}{n-1} \frac{N}{N-1} [ns_z^2 - \bar{z}(1-\bar{z}) + s_p^2]. \quad (11.5.2)$$

It is interesting to note that if  $\hat{\sigma}_\xi^2$  of (11.5.2) is divided by  $\hat{\sigma}_z^2 \equiv Ns_z^2/(N-1)$ , the result is  $\hat{\alpha}$ , the sample estimator of coefficient  $\alpha$  for binary items:

$$\begin{aligned} \frac{\hat{\sigma}_\xi^2}{\hat{\sigma}_z^2} &= \frac{n}{n-1} \left[ 1 - \frac{\bar{z}(1-\bar{z}) - s_p^2}{ns_z^2} \right] \\ &= \frac{n}{n-1} \left[ 1 - \frac{\sum_{g=1}^n p_g(1-p_g)}{s_x^2} \right]. \end{aligned} \quad (11.5.3)$$

The right-hand side of the first equality is essentially Tucker's formula for the Kuder-Richardson formula-20 sample reliability coefficient (Tucker, 1949,

Eq. 27). It is readily shown (see Exercise 4.21) that this is the same as the right-hand side of the second equality, which in turn is the same as the  $\alpha$  given by (9.4.5) for binary items.

It does *not* follow from this that  $\alpha$  should be considered a *generic* reliability coefficient. We have defined reliability as the squared correlation of observed score with true score. In the generic case, this correlation is *not* in general equal to the ratio of true-score variance to observed-score variance, as shown by (9.8.1). Such a variance ratio is inadequate for describing the relation of observed score to (generic) true score in situations where the (generic) error of measurement is biased and is correlated with (generic) true score.

**Higher-order moments.** Consider next the problem of estimating  $\mu_3$ , the third central moment (or cumulant) of the true scores. This problem will serve to illustrate the general method of using gsm's to find an unbiased estimate of a polynomial function of moments. The method consists of just two distinct stages:

1. In the first stage, the function of interest (in this case  $\mu_3$ ) is expressed as a (linear) function of gsm's. First of all,  $\mu_3$  is written in terms of noncentral moments, indicated by primes:

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'_1^3. \quad (11.5.4)$$

Next, each term on the right must be expressed as a linear function of gsm's. This can be done with the aid of the David, Kendall, and Barton (1966) tables. Where  $\bar{N} \rightarrow \infty$ , this is a simple matter for true-score moments since each term on the right of (11.5.4) is exactly equal to a population gsm. The desired gsm's for binary items can be found from a list in Lord (1960). From this list, we can write at once

$$\mu_3 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}' - 3 \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}' + 2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}',$$

where the primes denote population gsm's.

2. In the second stage, population gsm's are replaced by sample gsm's, simply by removing primes. Because of inheritance on the average, the result is the desired unbiased estimate. Thus *when  $\bar{N} \rightarrow \infty$ , an unbiased estimator for the third central moment (or cumulant) of the true score  $\xi$  for a binary-item test is*

$$\hat{\mu}_3 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} - 3 \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} + 2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (11.5.5)$$

For practical work, a third stage may be carried through: The unbiased estimate already obtained may be expressed in terms of more familiar sample statistics. Lord's Table 1 (1960) facilitates this step for binary items with

$\bar{N} \rightarrow \infty$ . Thus we write the following approximate computing formula, correct to terms of order  $1/N$  (we could write the exact formula if desired):

$$\hat{\mu}_3 = \frac{1}{n^{[3]}} [n^3 m_3 - 3n^2 s_z^2 + 6n^2 s(p_*, s_{u_g z}) + 6n^2 \bar{z} s_z^2 + 2n(2M'_3 - 3M'_2 + M'_1)], \quad (11.5.6)$$

where  $m_3$  is the third central sample moment of  $z$ ,  $M'_r$  is the  $r$ th sample moment about the origin of the sample item difficulties  $p_g$ , and  $s(p_*, s_{u_g z})$  is the sample covariance over items between  $p_g$  and  $s_{u_g z}$ ,  $s_{u_g z}$  being the sample covariance over people between the score  $u_{ga}$  on the binary item  $g$  and the relative score  $z_a$  on the entire test.

Similarly

$$\hat{k}_4 \equiv \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} - 4 \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} - 3 \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} + 12 \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} - 6 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (11.5.7)$$

is an unbiased estimator of the fourth cumulant of the true score. A similar relation holds for higher-order cumulants.

It is not necessary here to define and treat the bipolykays, which, the reader will remember, are linear functions of gsm's. Tables for converting bipolykays to gsm's and vice versa are given by Hooke. For the benefit of the reader who wishes to pursue the matter beyond the scope of this book, we note here two results that involve bipolykays.

The first result is that the unbiased estimate of a true-score cumulant given by (11.5.1), (11.5.5), or (11.5.7) is in each case itself a bipolykay. This is true of higher-order cumulants also.

The second result illustrates a type of formula obtainable. The sampling variance of  $\hat{\sigma}_{\xi}^2$  (see Eq. 11.5.1) for a sample taken from a doubly infinite population ( $\bar{n} = \bar{N} = \infty$ ) is

$$\begin{aligned} \text{Var } \hat{\sigma}_{\xi}^2 &= \frac{2}{N-1} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}' + \frac{1}{N} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}' + \frac{4}{n} \frac{N}{N-1} \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}' \\ &\quad + \frac{2}{n^{[2]}} \frac{N}{N-1} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}' + \frac{4}{n(N-1)} \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}' + \frac{4}{nN} \begin{pmatrix} 2 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}' \\ &\quad + \frac{2}{n^{[2]}(N-1)} \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}' + \frac{2}{n^{[2]}N} \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix}', \end{aligned} \quad (11.5.8)$$

where the arrays enclosed in large parentheses denote bipolykays, primes being

used to distinguish population values from sample values. Since bipolykays are inherited on the average, (11.5.8) without the primes would be an unbiased sample estimator of  $\text{Var } \hat{\sigma}_{\xi}^2$ . Hooke (1956a) has given general formulas.

The actual computation of quantities such as (11.5.8) is a complicated matter. Dayhoff (1966) has programmed high-speed computers both to derive formulas such as (11.5.8) and to evaluate them numerically.

### 11.6 Estimating the Relation of Observed Score to True Score

Consider the sample covariance between relative observed score  $z$  and relative true score  $\xi$ :

$$s_{z\xi} \equiv \frac{1}{N} \sum_a z_a \xi_a - \left( \frac{1}{N} \sum_a z_a \right) \left( \frac{1}{N} \sum_a \xi_a \right). \quad (11.6.1)$$

Take the expectation  $\mathcal{E}_I$  over all possible sets of  $n$ -item tests:

$$\begin{aligned} \mathcal{E}_I s_{z\xi} &= \frac{1}{N} \sum_a \xi_a \mathcal{E}_I z_a - \left( \frac{1}{N} \sum_a \mathcal{E}_I z_a \right) \left( \frac{1}{N} \sum_a \xi_a \right) \\ &= \frac{1}{N} \sum_a \xi_a^2 - \left( \frac{1}{N} \sum_a \xi_a \right)^2 = s_{\xi}^2, \end{aligned} \quad (11.6.2)$$

where  $s_{\xi}^2$  is the variance of  $\xi_a$  in the sample of  $N$  examinees. Finally take the expectation of (11.6.2) over the population of examinees:

$$\mathcal{E}_I \mathcal{E}_A s_{z\xi} = \mathcal{E}_A s_{\xi}^2 = \frac{N-1}{N} \sigma_{\xi}^2, \quad (11.6.3)$$

or

$$\mathcal{E}_I \mathcal{E}_A \frac{N}{N-1} s_{z\xi} = \sigma_{\xi}^2, \quad \text{or} \quad \mathcal{E}_I \sigma_{z\xi} = \sigma_{\xi}^2.$$

Thus, in matrix sampling, the quantity

$$\hat{\sigma}_{\xi}^2 \equiv \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

is not only an unbiased estimate over all matrix samples of  $\sigma_{\xi}^2$ , as in (11.5.1), but it is also an unbiased estimate for  $\mathcal{E}_I \sigma_{z\xi}$ , the expected value over all item samples of the covariance between observed score and true score.

Higher-order bivariate moments of the joint distribution of true score and observed score may be similarly estimated from higher-degree gsm's. The use and implications of such bivariate moments were illustrated and discussed at the end of Chapter 10.

### 11.7 Estimating the Relation between Scores on Parallel Test Forms

We shall call tests formed by random sampling of  $n$  items from the same population *randomly parallel* tests. An entire test theory can be built up for randomly parallel tests, but this theory would be essentially the same as item-sampling

theory. In such a theory, strong results are deduced from data on a single test form. We shall indicate only a few simple results in this chapter.

A derivation like the one used to obtain (11.6.3) shows that the expected value of the covariance between scores  $z$  and  $z'$  on two randomly parallel tests is

$$\mathcal{E}_I \mathcal{E}_A s_{zz'} = \frac{N - 1}{N} \sigma_z^2, \quad (11.7.1)$$

where  $\mathcal{E}_I$  indicates the expectation over all possible pairs of tests. Thus the expected value of  $s_{zz'}$  can be estimated from (11.5.1).

Higher-order bivariate moments of  $z$  and  $z'$  may be estimated from higher-degree bipolykays or gsm's. Lord (1960) has given some formulas for making such estimates for binary items when  $N$  is large.

## 11.8 Estimating the Observed-Score Statistics for Lengthened Tests

The same properties that make the gsm's and the bipolykays useful for estimating true-score moments also make them useful for estimating the effects of lengthening a test. Consider an infinite pool of test items from which a random sample of  $\bar{n}$  items is drawn. Now consider this sample as a finite pool of  $\bar{n}$  items and draw a random sample of  $n$  items from it. Administer the resulting  $n$ -item test to  $N$  examinees. Then any gsm computed for the data on the  $n$ -item test is an unbiased estimate of the same gsm for the  $\bar{n}$ -item test. Although this particular  $\bar{n}$ -item test contains the  $n$ -item test, this  $\bar{n}$ -item test was randomly chosen from the infinite pool; and consequently the estimator from the  $n$ -item test at hand can be used to estimate without bias the gsm of any other randomly chosen  $\bar{n}$ -item test. We often use the terms *estimate* and *unbiased* in this and later sections of this chapter in the broad senses to be understood from this reasoning; this allows us to "estimate" the statistics of a future sample from the statistics of a presently available sample.

Let us denote the relative score of examinee  $a$  on the  $\bar{n}$ -item test by

$$Z_a \equiv \frac{1}{\bar{n}} \sum_{G=1}^{\bar{n}} y_{Ga},$$

to distinguish this from  $z_a$ , the relative score on the  $n$ -item test. An estimate, unbiased in item sampling, for

$$s_Z^2 = \frac{1}{N} \sum_{a=1}^N Z_a^2 - \left( \frac{1}{N} \sum_{a=1}^N Z_a \right)^2$$

the (sample) variance of the scores of  $N$  examinees on an  $\bar{n}$ -item test, can be obtained from data on a short form of the test simply by replacing  $n$  by  $\bar{n}$  in (11.4.17). This convenient method of writing an unbiased estimate is a direct consequence of the fact that gsm's are inherited on the average. The  $N$  examinees may be thought of here as a finite population of  $N$ . The resulting formula,

valid when the  $n$  and the  $\bar{n}$  items are sampled from the same pool, is

$$\hat{s}_Z^2 \equiv \frac{(N-1)}{\bar{n}N} \left\{ -(\bar{n}-1) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (\bar{n}-1) \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\}, \quad (11.8.1)$$

where  $Z$  is the score on the  $\bar{n}$ -item test, and where the symmetric functions on the right are to be computed from data on the  $n$ -item test administered to  $N$  examinees.

If (11.4.5), (11.4.8), (11.4.11), and (11.4.1) are substituted in (11.8.1), we find that for tests composed of binary items, the sample variance of relative observed score  $s_Z^2$  for a random sample of  $N$  examinees on an  $\bar{n}$ -item test may be estimated without bias (in item sampling) from data for a random sample of  $N$  examinees on an  $n$ -item test by the formula

$$\begin{aligned} \hat{s}_Z^2 &\equiv \frac{1}{\bar{n}(n-1)} \{ n(\bar{n}-1)s_z^2 - (\bar{n}-n)[\bar{z}(1-\bar{z}) - s_p^2] \} \\ &\equiv \frac{1}{\bar{n}(n-1)} \left[ n(\bar{n}-1)s_z^2 + (\bar{n}-n) \frac{1}{n} \sum_{i=1}^n p_i q_i \right], \end{aligned} \quad (11.8.2)$$

where the statistics on the right are to be computed from the  $n$ -item test.

One can compare (11.8.2) with the classical formula (5.9.1b) for the variance of a lengthened test. The classical formula deals with population parameters for rigorously parallel tests; (11.8.2) deals with sample estimators for randomly parallel tests. The classical formula requires the correlation between parallel forms; (11.8.2) requires instead the mean and variance of the item scores. A detailed algebraic comparison of the formulas does not seem to yield any further useful insights.

Other parameters of lengthened tests may be estimated in similar fashion. Lord (1960, Table 2) gives some formulas.

## 11.9 Frequency Distribution of Errors of Measurement for Binary Items

If examinee  $a$  is a randomly selected examinee, then his observed score  $x_a$  on a particular random sample of  $n$  binary items may be considered to be the number of successes in  $n$  independent random trials. Thus in repeated testing with different random samples of  $n$  items each, the frequency distribution of the observed score of examinee  $a$  will be

$$f(x_a) = \binom{n}{x_a} \xi_a^{x_a} (1 - \xi_a)^{n-x_a}. \quad (11.9.1)$$

This is an ordinary binomial distribution with probability of success  $\xi_a$ . An illustration of this relation between observed score and true score appears as Fig. 23.2.1.

If the error of measurement is defined as  $\eta_a \equiv x_a - n\xi_a$ , the frequency distribution of  $\eta_a$  is the same as (11.9.1) except for a change of origin, the new origin being at the mean of the distribution of  $\eta_a$ .

It is well known that  $x_a$  is a sufficient statistic for estimating  $\xi_a$ . Since  $x_a$  has a known frequency distribution, the item-sampling model, when applicable, provides much more information about the errors of measurement than does classical test theory. In particular, it is clear that for tests composed of binary items:

1. The expected error for a given examinee over repeated testings is zero.
2. *The variance of these errors of measurement for examinee  $a$  over repeated testings, denoted by  $\sigma_I^2(\eta_a)$ , is given by the usual binomial formula*

$$\sigma_I^2(\eta_a) = n\xi_a(1 - \xi_a). \quad (11.9.2)$$

This is largest when  $\xi_a = \frac{1}{2}$ , decreasing sharply for  $\xi_a$  near zero or one. *An estimate, unbiased in item sampling, of this error variance for a given examinee  $a$  is*

$$\hat{\sigma}_I^2(\eta_a) \equiv \frac{1}{n-1} x_a(n - x_a). \quad (11.9.3)$$

If (11.9.3) is averaged over all examinees, *the estimate, unbiased in matrix sampling, for the group error variance is found to be*

$$\hat{\sigma}_I^2(\eta_s) \equiv \frac{1}{n-1} [\bar{x}(n - \bar{x}) - s_x^2]. \quad (11.9.4)$$

This estimate can be computed directly from the mean and variance of the observed test scores.

3. If  $\xi_a < \frac{1}{2}$ , the distribution over randomly parallel test forms of the errors of measurement for examinee  $a$  is positively skewed; if  $\xi_a = \frac{1}{2}$ , it is symmetric; if  $\xi_a > \frac{1}{2}$ , it is negatively skewed. Clearly *the distribution over randomly parallel forms of the errors of measurement is not independent of true score*.

Equation (11.9.2) may seem paradoxical from two points of view:

1. According to the equation, the standard error of measurement of a test for a particular examinee is determined entirely by the number of items in the test and by his true score. The examinee's true score depends on how difficult the items are for him, but not on other properties of the items unrelated to their difficulty. Since the standard error of measurement does not depend on the other properties of the test items, why devote great efforts (see Sections 15.4 and 15.10) to developing and selecting highly discriminating\* items when building a test?

---

\* A discriminating item is one that correlates highly with some criterion.

2. The standard error of measurement given by (11.9.2) is smallest for examinees whose true scores are nearest one or zero. Should it not follow from this that the best measuring instrument is a test composed of items so easy that everyone will have a relative true score near one, or so hard that everyone will have a relative true score near zero?

The answer to both these questions is that the effectiveness of a test as a measuring instrument usually does not depend merely on the standard error of measurement, but rather on the ratio of the standard error of measurement to the standard deviation of observed scores in the group. The more discriminating the test items, the larger will be the standard deviation of observed scores, other things being equal; and hence, the less will be the danger that true differences will be swamped by random errors of measurement and lost to view.

The small standard errors of measurement that result when a test is made very easy are not beneficial because the standard deviation of observed scores for such tests is also small. This is most apparent in the limiting case when the test is so easy (or so difficult) that everyone gets a perfect (or a zero) score and both standard deviations are zero. Even though in this case there are no errors of measurement at all, such a test obviously is not discriminating among examinees and thus is not a useful measuring instrument.

### **11.10 Item Sampling as a Technique in Research Design**

Item sampling can be used not only to provide a theoretical basis for mental test theory; sometimes it also may be introduced advantageously into the design of a research study. There are at least three obvious considerations:

1. If only a limited amount of time can be demanded of each research subject, the total amount of information obtained from a given number of subjects may be greatly increased by item sampling.
2. If a test can be administered to only one examinee at a time, the examiner's time may be the limiting factor; more information about a group of examinees may be obtained by giving a few items to each examinee than by giving the entire test to just a few examinees.
3. With certain tests, scoring costs may be the limiting factor; in this case, it would be better to score a few items from the answer sheets of every examinee than to score every item on the answer sheets of a few examinees.

Any use of item sampling obviously assumes that performance on an item does not depend on the context in which the item occurs. This assumption is commonly made in test theory and in practical work with item analysis data. It is supported in a rough way by various empirical studies (e.g., French and Greer, 1964; Sax and Cromack, 1966). However, the assumption certainly should not be made without caution. In particular, item sampling is clearly inappropriate

if the purpose is to estimate performance on a preexisting speeded or partially speeded test, since the examinee's performance on an item in a speeded test depends very much on whether he has time to reach it.

Problems of the relative efficiency of different experimental designs involving item sampling constitute an area in which little work has been done. This is especially true when variances, as well as means, of scores are to be estimated. A few sampling variances for estimated score means will be given here, for the simplest cases only.

We assume throughout that the responses to an item are independent of the context in which it is administered. We shall denote the relative score of examinee  $a$  on  $n$  items by  $z_a$ ; on all  $\bar{n}$  items, by  $Z_a$  for finite  $\bar{n}$ , and by  $\xi_a$  for infinite  $\bar{n}$ . We shall denote the difficulty of item  $i$  for  $N$  examinees by  $p_i$ ; for all  $\bar{N}$  examinees, by  $P_i$  for finite  $\bar{N}$ , and by  $\pi_i$  for infinite  $\bar{N}$ .

### 11.11 Estimating a Mean from a Single Item Sample

If a single random sample of  $n$  items is drawn without replacement from a population of  $\bar{n}$  binary items and administered to each of  $N$  examinees, then it is clear that for the given group of examinees, their mean relative score  $\bar{z} \equiv \sum^N z_a/N$  on the  $n$ -item test is an unbiased estimator of their mean relative score  $\bar{Z} \equiv \sum^N Z_a/N$  on the  $\bar{n}$ -item test.

The sampling variance of  $\bar{z}$  due to sampling of examinees from an infinite population is well known to be

$$\text{Var}_A(\bar{z}) = \sigma_z^2/N, \quad (11.11.1)$$

where  $\sigma_z^2$  is the population variance of  $z$ . If the population of examinees is finite and (as throughout this chapter) sampling is without replacement, then

$$\text{Var}_A(\bar{z}) = \frac{(\bar{N} - N)}{(\bar{N} - 1)} \frac{\sigma_z^2}{N}. \quad (11.11.2)$$

If items (not examinees) are sampled from a possibly finite population of  $\bar{n}$  items, then the sampling variance of  $\bar{z} = \bar{p} \equiv \sum^n p_i/n$  over all possible  $n$ -item tests is, by symmetry,

$$\text{Var}_I(\bar{z}) = \frac{(\bar{n} - n)}{(\bar{n} - 1)} \frac{\sigma_p^2}{n}, \quad (11.11.3)$$

where  $\sigma_p^2$  is the variance of the  $\bar{n}$  item difficulties in the population of items. An unbiased estimate of this sampling variance is

$$\hat{\text{Var}}_I(\bar{z}) = \frac{(\bar{n} - n)}{\bar{n}} \frac{s_p^2}{n - 1}, \quad (11.11.4)$$

where  $s_p^2$  is the variance of the  $n$  item difficulties in the sample of items.

If a random matrix sample of  $n$  items and  $N$  examinees is drawn from a population of  $\bar{n}$  items and  $\bar{N}$  examinees,  $\bar{z}$  will still be an unbiased estimate of  $\bar{Z}$ . Hooke (1956a) has given the sampling variance of this estimate in terms of bipolykays. It is presented in these terms here for illustrative purposes:

$$\text{Var}_{IA}[1] = \left( \frac{1}{nN} - \frac{1}{\bar{n}\bar{N}} \right) \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}' + \left( \frac{1}{n} - \frac{1}{\bar{n}} \right) \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}' + \left( \frac{1}{N} - \frac{1}{\bar{N}} \right) \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}'. \quad (11.11.5)$$

In terms of more familiar test parameters, the sampling variance of the estimate for binary items is

$$\begin{aligned} \text{Var}_{IA}(\bar{z}) &= \frac{1}{nN(\bar{n}-1)(\bar{N}-1)} [\bar{N}(N-1)(\bar{n}-n)\sigma_P^2 \\ &\quad + \bar{n}(n-1)(\bar{N}-N)\sigma_Z^2 + (\bar{n}-n)(\bar{N}-N)\bar{Z}(1-\bar{Z})]. \end{aligned} \quad (11.11.6)$$

We can obtain a clearer understanding of (11.11.6) by allowing  $\bar{n} \rightarrow \infty$  and  $\bar{N} \rightarrow \infty$ . Then, substituting  $\xi$  for  $Z$  and  $\pi$  for  $P$ , we have

$$\text{Var}_{IA}(\bar{z}) = \left( 1 - \frac{1}{N} \right) \frac{\sigma_\pi^2}{n} + \left( 1 - \frac{1}{n} \right) \frac{\sigma_\xi^2}{N} + \frac{\mu(1-\mu)}{nN}. \quad (11.11.7)$$

Suppose that  $nN$ , the total number of observations  $y_{ga}$ , is fixed. Then, if  $\sigma_\pi^2$  is small enough compared to  $\sigma_\xi^2$ , it will be advantageous to reduce  $n$  while increasing  $N$  proportionately, so long as  $n \geq 1$ .

If the efficiency of the experimental design depends on (11.11.7), then the type of item sampling under discussion will be unprofitable in the typical situation where  $N > n$  and where  $\sigma_\pi^2$  is not small compared to  $\sigma_\xi^2$ . If  $\sigma_\pi^2$  is large enough compared to  $\sigma_\xi^2$ , it will be advantageous to reduce  $N$  while increasing  $n$  proportionately, so long as  $N \geq 1$ . It can be shown by standard minimization procedures that when  $nN$  is fixed, (11.11.6) is minimized by choosing  $n$  so that

$$\frac{n}{N} = \frac{\bar{n}[\bar{N}\sigma_P^2 + \sigma_Z^2 - \bar{Z}(1-\bar{Z})]}{\bar{N}[\sigma_P^2 + \bar{n}\sigma_Z^2 - \bar{Z}(1-\bar{Z})]}. \quad (11.11.8)$$

The purpose of writing the foregoing formulas is to show how the efficiency of the estimate of the mean depends on the number of items and on the number of examinees. If in practice we wish to estimate the sampling variance of  $\bar{z}$ , we can estimate the right-hand side of (11.11.6) by first estimating  $\sigma_Z^2$  from (11.8.2) and then estimating  $\sigma_P^2$  from a similar formula, obtained from (11.8.2) by substituting  $P$  for  $Z$ ,  $p$  for  $z$ ,  $z$  for  $p$ , and  $N$  for  $n$ . Alternatively, and perhaps more conveniently, we could estimate the sampling variance of this  $\bar{z}$  by applying an ordinary variance-components analysis to the raw data.

### 11.12 Estimating a Mean by Multiple Matrix Sampling

The methods of the preceding section do not make full use of the advantages of item sampling. Under a more efficient procedure to be outlined in this section, the examiner administers different samples of binary items to different subgroups of examinees.\* This procedure draws on a mathematical formulation that has arisen from certain unpublished suggestions of Dr. William W. Turnbull.

Suppose  $M$  nonoverlapping random samples of  $n$  binary items each are drawn (without replacement) from an  $\bar{n}$ -item test and treated as separate subtests; it is not required that  $M = \bar{n}/n$  or that  $M = \bar{N}/N$ . A different subtest is administered to each of  $M$  nonoverlapping random samples of  $N$  examinees drawn from a population of  $\bar{N}$  examinees. If  $\bar{z}_m$  is the mean relative score of subgroup  $m$  on an  $n$ -item test, then the average  $\bar{z}_m$  is an unbiased estimator of  $\bar{Z}$ , the mean score of the  $\bar{N}$  examinees on the  $\bar{n}$ -item test. In symbols,

$$\hat{\bar{Z}} \equiv \frac{1}{M} \sum_{m=1}^M \bar{z}_m. \quad (11.12.1)$$

By the formula for the variance of a sum,

$$\begin{aligned} \text{Var}_{IA}(\hat{\bar{Z}}) &= \frac{1}{M^2} \left[ \sum_{m=1}^M \text{Var}_{IA}(\bar{z}_m) + \sum_{m \neq m'} \text{Cov}_{IA}(\bar{z}_m, \bar{z}_{m'}) \right] \\ &= \frac{1}{M} \text{Var}_{IA}(\bar{z}_m) + \frac{M-1}{M} \text{Cov}_{IA}(\bar{z}_m, \bar{z}_{m'}). \end{aligned}$$

It may be shown that

$$\text{Cov}_{IA}(\bar{z}_m, \bar{z}_{m'}) = \frac{1}{(\bar{n}-1)(\bar{N}-1)} [\bar{n}\sigma_Z^2 + \bar{N}\sigma_P^2 - \bar{Z}(1-\bar{Z})]. \quad (11.12.2)$$

From (11.11.6) and (11.12.2), we finally find that for binary items,

$$\begin{aligned} \text{Var}_{IA}(\hat{\bar{Z}}) &= \frac{1}{nNM(\bar{n}-1)(\bar{N}-1)} \\ &\times \{ \bar{N}\sigma_P^2[(\bar{n}-n)(\bar{N}-1) - nN(M-1)] \\ &\quad + \bar{n}\sigma_Z^2[(\bar{N}-N)(n-1) - nN(M-1)] \\ &\quad + \bar{Z}(1-\bar{Z})[(\bar{n}-n)(\bar{N}-N) + nN(M-1)] \}. \end{aligned} \quad (11.12.3)$$

In the special case where  $M = \bar{n}/n = \bar{N}/N$ , this reduces to

$$\text{Var}_{IA}(\hat{\bar{Z}}) = \frac{M-1}{(\bar{n}-1)(\bar{N}-1)} [\bar{Z}(1-\bar{Z}) - \sigma_P^2 - \sigma_Z^2]. \quad (11.12.4)$$

---

\* The reader uninterested in such matters of experimental design should omit the remainder of this chapter.

For the practical worker using number-right scores denoted by  $x_a \equiv nz_a$  and  $X_a \equiv \bar{n}Z_a$ , we note that the corresponding sampling variance of

$$\hat{X} \equiv \bar{n}\hat{Z}$$

is

$$\text{Var}_{IA}(\hat{X}) = \frac{M - 1}{(\bar{n} - 1)(\bar{N} - 1)} [\bar{X}(\bar{n} - \bar{X}) - \bar{n}^2\sigma_P^2 - \sigma_X^2].$$

If  $\bar{N} \rightarrow \infty$  and  $M = \bar{n}/n$ , then, writing  $\hat{\mu}$  for  $\hat{Z}$  and  $\pi$  for  $P$ , we find that (11.12.3) becomes

$$\text{Var}_{IA}(\hat{\mu}) = \frac{1}{N\bar{n}(\bar{n} - 1)} \{ \bar{n}(n - 1)\sigma_Z^2 + (\bar{n} - n)[\bar{Z}(1 - \bar{Z}) - \sigma_\pi^2] \}. \quad (11.12.5)$$

If (11.12.5) is written

$$\text{Var}_{IA}(\hat{\mu}) = \frac{1}{\bar{n}N(\bar{n} - 1)} \{ \bar{n}[\bar{Z}(1 - \bar{Z}) - \sigma_\pi^2 - \sigma_Z^2] + n[\bar{n}\sigma_Z^2 - \bar{Z}(1 - \bar{Z}) + \sigma_\pi^2] \},$$

the last term in brackets is equal to  $(\bar{n} - 1)\sigma_Z^2 r_{20}$ , where  $r_{20}$  is the Kuder-Richardson reliability (see 11.5.3) of the  $\bar{n}$ -item test. We see from this formulation that if  $n$  is decreased and  $N$  is increased while  $nN$  is held constant, (11.12.5) will decrease so long as  $n > 1$ , provided that  $r_{20}$  is positive. Although tests with negative  $r_{20}$  are possible (for example, when  $\sigma_Z^2 = 0$ ), they are ordinarily of no practical interest.

*Thus, assuming that performance on an item is independent of the context in which the item appears, the mean performance of a very large group of examinees  $\bar{N} \rightarrow \infty$  on a pool of  $\bar{n}$  binary items can best be estimated by administering a different item to each of  $\bar{n}$  nonoverlapping random samples of  $N$  examinees. This is better than administering all  $\bar{n}$  items to a single sample of  $N$  examinees. It is better than using any other split of the item pool into nonoverlapping subtests, each administered to a different sample of  $N$  examinees.*

For example, suppose  $\bar{n} = 36$ ,  $\sigma_Z^2 = 0.02$ ,  $\sigma_P^2 = 0.05$ , and  $\bar{Z} = 0.5$ . If we administer all 36 items to 252 examinees, the standard error of  $\bar{n}\hat{\mu}$  is simply  $\bar{n}\sigma_Z/\sqrt{N} = 0.32$ . If we administer each item to a different group of  $N = 252$  examinees and let  $\hat{\mu}$  be the average of the 36 different item difficulties so obtained, then the standard error of  $\bar{n}\hat{\mu}$  is, by (11.12.5), only 0.17. To obtain such a reduction in sampling error without item sampling would require quadrupling the number of examinees to be tested with the entire pool of 36 items.

One further conclusion may be suggested. Especially if  $MN$ , the total number of examinees tested, is large, as when test norms are being developed, it is important that every item be administered, unless  $\bar{n}$  is very large. Consider the following numerical results obtained from (11.12.4) for  $\bar{n} = 36$ ,  $\sigma_P^2 = 0.05$ ,  $\bar{Z} = 0.5$ ,  $\sigma_Z^2 = 0.02$ , and  $\bar{N} = MN = 25,200$ . If the 36 items are

divided into  $M = 6$  subtests of 6 items each and each subtest is given to  $25,200/6 = 4200$  examinees, then the standard error of  $\hat{X} = \bar{n}\hat{Z}$  is 0.036. If they are divided to make  $M = 7$  subtests of 5 items each, one item being left out altogether, and each subtest is given to  $25,200/7 = 3600$  examinees, the standard error is 0.234. If they are divided to make  $M = 6$  subtests of 5 items each and each subtest is given to  $25,200/6 = 4200$  examinees, the standard error is 0.610.

The drastic increase in the last two standard errors over the first is due to the failure to administer all 36 items. Omitting even one item has a very bad effect. The smaller  $\bar{n}$  is, the worse the effect.

Until now, we have assumed that each item is administered to the same number of examinees, and that the estimate of  $\bar{Z}$  is to be an *unweighted* sum of examinees' scores. If the number of examinees taking each particular item were chosen optimally in relation to the item variance (or even approximately so), and if a correspondingly *weighted* sum of examinees' scores were used to estimate  $\bar{Z}$ , then the sampling variance of the weighted estimate could ordinarily be reduced below the sampling variance of the unweighted estimate. Relevant formulas for such weighted estimates and for their sampling variances are not presently available.

The item-sampling procedure is recommended primarily for a situation such as one of the following:

1. It is impossible or impractical to administer all items to every examinee.
2. The attempt to test all examinees fully would itself introduce some bias into the results because of poor cooperation from examinees or because of selective loss of examinees.
3. There is no preexisting test on which performance must be estimated, and the research worker is therefore free to administer different items to different examinees as he pleases within a fixed testing period.

In such situations, appropriate item-sampling procedures may be really worthwhile. In other situations, the reduction in sampling error may not justify the risk taken in assuming that item response is independent of item context.

Item sampling has been successfully used on a large scale in the National Longitudinal Study of Mathematical Abilities. A preliminary report of certain aspects of this work was given by Cahen (1967).

### 11.13 Estimating Group Mean Differences

When we wish to estimate a group mean score  $\bar{Z}$  by sampling from a moderate-sized pool of  $\bar{n}$  items, it is important that we utilize every item, as pointed out above. The reason is that otherwise the psychological dimension measured by  $\hat{Z}$  will not coincide with the dimension measured by  $\bar{Z}$ . This consideration loses most of its importance if the investigator's purpose is to study differences between groups (or between experimental treatments).

In such a case the investigator must estimate  $\bar{Z}' - \bar{Z}''$ , where the primes are used to distinguish two different populations of examinees. In one common design for estimating  $\bar{Z}' - \bar{Z}''$ , the same  $n$  items are administered to both groups of examinees. This design ensures that observed differences between  $\bar{z}'$  and  $\bar{z}''$  cannot be due to differences between items; hence it is no longer so important that the sample of items contain all or most of the items in the available pool.

Suppose we draw a single random sample of  $n$  items (as in Section 11.11) and administer it to two random samples of  $N$  examinees. The difference  $\bar{z}' - \bar{z}''$  between samples in mean observed score is then to be used as an unbiased estimate of the difference  $\bar{Z}' - \bar{Z}''$  between populations in mean true score.

Consider the sampling variance of the estimate  $\bar{z}' - \bar{z}''$ . As usual,

$$\text{Var}_{IA}(\bar{z}' - \bar{z}'') = \text{Var}_{IA}(\bar{z}') + \text{Var}_{IA}(\bar{z}'') - 2 \text{Cov}_{IA}(\bar{z}', \bar{z}'').$$

It can be shown that

$$\text{Cov}_{IA}(\bar{z}', \bar{z}'') = \frac{\bar{n} - n}{n(\bar{n} - 1)} \text{Cov}_I(P'_I, P''_I), \quad (11.13.1)$$

the covariance on the right being the covariance, over all  $\bar{n}$  items, between the item difficulties for one population of examinees and the item difficulties for the other population.

In the special case where the two populations do not differ from each other,

$$\text{Cov}_{IA}(\bar{z}', \bar{z}'') = \frac{\bar{n} - n}{n(\bar{n} - 1)} \sigma_P^2. \quad (11.13.2)$$

In the special case where the two populations are the same and  $N' = N''$ , we find from (11.13.1), (11.13.2), and (11.11.6) that

$$\begin{aligned} \text{Var}_{IA}(\bar{z}' - \bar{z}'') &= \frac{2(\bar{N} - N)}{nN(\bar{n} - 1)(\bar{N} - 1)} \\ &\times \{\bar{n}(n - 1)\sigma_Z^2 + (\bar{n} - n)[\bar{Z}(1 - \bar{Z}) - \sigma_P^2]\}. \end{aligned} \quad (11.13.3)$$

If the two populations are infinite, as is commonly the case, then

$$\begin{aligned} \text{Var}_{IA}(\bar{z}' - \bar{z}'') &= \frac{2}{nN(\bar{n} - 1)} \\ &\times \{\bar{n}(n - 1)\sigma_Z^2 + (\bar{n} - n)[\bar{Z}(1 - \bar{Z}) - \sigma_\pi^2]\}. \end{aligned} \quad (11.13.4)$$

If (11.13.4) is written

$$\begin{aligned} \text{Var}_{IA}(\bar{z}' - \bar{z}'') &= \frac{2}{nN(\bar{n} - 1)} \{\bar{n}[\bar{Z}(1 - \bar{Z}) - \sigma_\pi^2 - \sigma_Z^2] + n[\bar{n}\sigma_Z^2 - \bar{Z}(1 - \bar{Z}) + \sigma_\pi^2]\}, \end{aligned} \quad (11.13.5)$$

the last term in brackets is equal to  $(\bar{n} - 1)\sigma_Z^2 r_{20}$ , where  $r_{20}$  is the Kuder-Richardson reliability of the  $\bar{n}$ -item test. We see from this formulation that if

$n$  is decreased and  $N$  is increased while  $nN$  is held constant, (11.13.5) will decrease so long as  $n \geq 1$ , provided always that  $r_{20}$  is positive.

The conclusion to be drawn is that whenever the two populations are sufficiently similar [especially when  $\text{Cov}_I(\pi'_I, \pi''_I)$  is not too different from  $\sigma_\pi^2$ ], then designs of the type described will be profitable. A research worker who can increase the number of examinees proportionately as he cuts down on testing time may find this type of item sampling of benefit in situations that call for comparison of group means.

Again we should point out that the primary purpose in giving the foregoing formulas is to display the effect of variations in  $n$  and  $N$  on the efficiency of the experimental design. If a practical worker wishes to compare a computed value of  $\bar{z}' - \bar{z}''$  with its sampling error under a null hypothesis, he will find it most convenient to carry out a straightforward analysis of variance components on his data.

### 11.14 Estimating Observed-Score Variances by Item Sampling

If a sample of  $n$  items is administered to a sample of  $N$  examinees, the gsm's of the sample can be used to obtain an estimate of  $\sigma_Z^2$  that is unbiased in matrix sampling. We find that

$$\hat{\sigma}_Z^2 \equiv \frac{N(\bar{N} - 1)}{\bar{n}(n - 1)\bar{N}(N - 1)} \{n(\bar{n} - 1)s_z^2 - (\bar{n} - n)[\bar{z}(1 - \bar{z}) - s_p^2]\} \quad (11.14.1)$$

is such an estimate. Except for the term  $N(\bar{N} - 1)/\bar{N}(N - 1)$ , this is the same as the estimate for the variance of a lengthened test given in (11.8.5).

The sampling variance of  $\hat{\sigma}_Z^2$  could be written in terms of fourth-degree bipolykays from formulas given by Hooke. The task of estimating the sampling variance from actual sets of data with sizable  $n$  and  $N$  would not be a casual undertaking, however.

As we saw in Section 11.12, it is much better to administer many different  $n$ -item subtests than just one. If this is done, it is again important (assuming special weighting procedures are not used) that every item appear an equal number of times; that all  $(\bar{N})$  pairs of items appear in the subtests, if possible; and that each pair be administered to the same number of examinees. When all  $(\bar{N})$  pairs cannot be used, good balanced designs may sometimes be found with the help of tables of balanced incomplete blocks (for example, Fisher and Yates, 1938, Tables 17–19; also see references in Wilks, 1962, Section 8.6d). A more detailed discussion is given by Knapp (in press).

### References and Selected Readings

- BEHNKEN, D. W., Sampling moments of means from finite multivariate populations. *Annals of Mathematical Statistics*, 1961, **32**, 406–413.
- CAHEN, L. S., The estimation of mean achievement scores for schools by the item-sampling technique. Paper presented at the meeting of the Psychometric Society, Madison, March, 1967.

- CORNFIELD, J., and J. W. TUKEY, Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, **27**, 907-949.
- CRONBACH, L. J., P. SCHÖNEMANN, and D. MCKIE, Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 1965, **25**, 291-312.
- DAVID, F. N., M. G. KENDALL, and D. E. BARTON, Symmetric function and allied tables. London: Cambridge University Press, 1966.
- DAYHOFF, E., Generalized polykays, an extension of simple polykays and bipolykays. *Annals of Mathematical Statistics*, 1966, **37**, 226-241.
- FISHER, R. A., and F. YATES, *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver & Boyd, 1938.
- FRENCH, J. L., and D. GREER, Effect of test-item arrangement on physiological and psychological behavior in primary-school children. *Journal of Educational Measurement*, 1964, **1**, 151-153.
- HOOKE, R., Symmetric functions of a two-way array. *Annals of Mathematical Statistics*, 1956, **27**, 55-79. (a)
- HOOKE, R., Some applications of bipolykays to the estimation of variance components and their moments. *Annals of Mathematical Statistics*, 1956, **27**, 80-98. (b)
- KNAPP, T. R., An application of balanced incomplete block designs to the estimation of test norms. *Educational and Psychological Measurement*, in press.
- LOEVINGER, JANE, Person and population as psychometric concepts. *Psychological Review*, 1965, **72**, 143-155.
- LORD, F. M., Use of true-score theory to predict moments of univariate and bivariate observed-score distributions. *Psychometrika*, 1960, **25**, 325-342.
- MCKIE, D., A model for the generalizability of tests constructed by stratified sampling. Urbana, Ill.: University of Illinois, 1965. Doctoral dissertation.
- OSBURN, H. G., A note on design of test experiments. *Educational and Psychological Measurement*, in press.
- RAJARATNAM, NAGESWARI, L. J. CRONBACH, and GOLDINE C. GLESER, Generalizability of stratified-parallel tests. *Psychometrika*, 1965, **30**, 39-56.
- ROBSON, D. S., Applications of multivariate polykays to the theory of unbiased ratio-type estimation. *Journal of the American Statistical Association*, 1957, **52**, 511-522.
- SAX, G., and T. R. CROMACK, The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement*, 1966, **3**, 309-311.
- SHOEMAKER, D. M., An empirical study of generalizability coefficients for matched and unmatched data. University of Houston, 1966. Doctoral dissertation.
- TUCKER, L. R., A note on the estimation of test reliability by the Kuder-Richardson formula (20). *Psychometrika*, 1949, **14**, 117-119.
- WILKS, S. S., *Mathematical statistics*. New York: Wiley, 1962.

**Part 4**

**VALIDITY AND**

**TEST CONSTRUCTION THEORY**



## CHAPTER 12

# VALIDITY

### 12.1 Introduction

Semantically, and in relatively inexplicit language, we may define the validity of a test as the extent to which it measures or predicts some criterion of interest. This criterion may be some theoretical construct, such as “mathematical ability”, or it may be some future performance, such as success in a test theory course. Typically a test designed to measure a psychological construct is used to predict many performance criteria. Hence any test has many such validities, and these validities are the meaningful measures of the value of that test. These validities also give evidence as to the value of the underlying construct.

The most common way of explicating the general concept of validity in statistical terms is as a *validity coefficient*, the (absolute value of the) linear product moment correlation between the test and a specified criterion. However, this is by no means the only possible technical definition that can be used for this general concept. Indeed, as we shall show in this chapter and note again in later chapters, this definition is very specific in nature. For many purposes it is very useful. For many other purposes, however, other technical definitions, such as the definition of the discriminating power of a test (Chapter 16), are more valuable.

Regardless of the particular statistical explication of the concept of validity, it will be fruitful for our purposes to distinguish two general kinds of validity. We refer to the first and more common as an *empirical validity*, defining it as the degree of association between the measurement and some other observable measurement. When the general concept of validity is explicated as a validity coefficient, it is the correlation of the given observable measurement with a second *observable* variable. As such, it is a measure of the linear predictive ability of the test with respect to the observable criterion. Most validity problems in testing technology have been approached as problems in linear prediction of observable quantities.

We refer to the second and perhaps more broadly meaningful kind of validity as a *theoretical validity*. The validity coefficient is then the correlation of an observed variable with some *theoretical construct* (latent variable) of interest. The present chapter is primarily concerned with empirical validity, although in Section 12.11 we briefly discuss the important case of theoretical validity called

*construct validity.* One theoretical validity, which is not necessarily a construct validity, is the square root of the reliability of a test.

The suggestion that the validity coefficient of one measurement with respect to a second measurement, defined as the correlation between them, can be interpreted as a measure of the degree to which the first is a valid linear predictor of the second can be justified within the theory of linear regression functions. This theory is developed for the one- and two-variable cases in the next two sections and then, in the following section, the related theories of multiple and partial correlation are presented. In Section 12.5, a matrix-algebra presentation of correlation and regression theories for the  $n$ -variable case is given. Our approach to these theories is essentially that taken by Cramér (1946) and Wilks (1963). In some respects these sources are not sufficiently complete to serve the specific needs of psychologists interested in undertaking validity studies. However, since these theories are standard topics in general statistical theory, we shall omit the details of proofs, each of which can be found in Cramér (1946), Wilks (1963), Kendall and Stuart (1961), or Anderson (1948). Efficient computational techniques for employing high-speed computers in the application of these methods are given by Beaton (1964) in terms of special matrix operators. The most important of these, the *sweep operator*, is particularly useful for the regression methods presented in this and the following chapter. In this chapter, we shall be concerned only with the mathematical theory of regression and correlation. We defer questions of sampling and inference to the next chapter.

## 12.2 Regression and Prediction

If we desire to predict a future value of  $X_0$  on the basis of a function of an observation  $x_1$  of a variable  $X_1$ , a reasonable measure of the imprecision of the prediction (for a specified value  $x_1$ ) is the conditional expected value of the square of the error  $X_0 - u(x_1)$ , where  $u$  is the function of  $x_1$  used to predict  $X_0$ . If no restrictions are placed on the form of the function  $u$ , then, with  $\mathcal{E}$  denoting expectation with respect to  $X_0$ ,

$$\mathcal{E}\{[X_0 - u(x_1)]^2 | x_1\} \quad (12.2.1)$$

is minimized when  $u(x_1) = \mu(X_0 | x_1)$ , the conditional expectation of  $X_0$ , given  $x_1$ . The function  $\mu$  is referred to as the *regression function* of  $X_0$  and  $x_1$ . The minimum value of the squared error is  $\sigma^2(X_0 | x_1)$ , the conditional variance of  $X_0$ , given  $x_1$ . The quantity

$$1 - \frac{\sigma^2(X_0 | x_1)}{\sigma^2(X_0)} \quad (12.2.1a)$$

may be taken as a natural measure of the predictability of the random variable  $X_0$ , given that  $X_1 = x_1$ . This quantity, of course, depends on  $x_1$ . However, since (12.2.1) is thus minimized for every value of  $x_1$ , this procedure will minimize  $\mathcal{E}[X_0 - u(X_1)]^2$ , where expectation is taken over both  $X_0$  and  $X_1$ . A

measure of the average predictability of  $X_0$  from  $X_1$  is given by

$$\mathcal{E}\left[1 - \frac{\sigma^2(X_0 | x_1)}{\sigma^2(X_0)}\right], \quad (12.2.1b)$$

where expectation is taken with respect to  $X_0$  and  $X_1$ . This quantity, called the *correlation ratio* of  $X_0$  with respect to  $X_1$ , may be expressed as

$$\eta_{0 \cdot 1}^2 = 1 - \frac{\mathcal{E}[\sigma^2(X_0 | x_1)]}{\sigma^2(X_0)} = \frac{\sigma^2[\mathcal{E}(X_0 | x_1)]}{\sigma^2(X_0)}, \quad (12.2.2)$$

where the second expression for  $\eta_{0 \cdot 1}^2$  is obtained from the first by using the basic relation

$$\sigma^2(X_0) = \mathcal{E}[\sigma^2(X_0 | x_1)] + \sigma^2[\mathcal{E}(X_0 | x_1)].$$

This theory can be extended to the case where values of a number of random variables  $X_1, X_2, \dots, X_n$  are used to predict a value for the random variable  $X_0$ . A completely analogous development leads to the choice of the function  $u(x_1, x_2, \dots, x_n) = \mu(X_0 | x_1, x_2, \dots, x_n)$ , the *regression function of  $X_0$  on  $x_1, x_2, \dots, x_n$* . The analogous measure of average predictability is

$$1 - \frac{\mathcal{E}[\sigma^2(X_0 | x_1, x_2, \dots, x_n)]}{\sigma^2(X_0)} = \frac{\sigma^2[\mathcal{E}(X_0 | x_1, x_2, \dots, x_n)]}{\sigma^2(X_0)}, \quad (12.2.3)$$

the *multiple correlation ratio* of  $X_0$  with respect to  $X_1, X_2, \dots, X_n$ .

For some distributions, of which the multivariate normal is the most important example, all such regression functions are linear functions. Although these regression functions will not be linear in more general contexts, we may still wish to restrict ourselves to linear functions of  $x_1, x_2, \dots, x_n$  in predicting  $X_0$ . In determining a test score from a set of item scores, we almost always wish to restrict ourselves to a linear function of the item scores. Similarly, in attempting to predict a criterion from a group of predictor test variables, we usually restrict ourselves to linear prediction methods.

There are several justifications for this restriction. The first is computational simplicity. Linear methods are easier to apply than methods involving true regression functions, and no assumption about the nature of the true regression function is required. Also it is possible to apply linear methods and yet consider certain kinds of nonlinear prediction functions: For example, the linear prediction function  $\alpha + \beta_1 x_1 + \beta_2 x_2$  can be generalized to include quadratic terms  $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$ . This latter function, however, is linear in the coefficients and hence can be handled by linear prediction methods. Also it is often possible to linearize relationships by employing suitable transformations: For example, if  $\mathcal{E}(y) = a + b e^{-x}$ ,  $x > 0$ , then  $\mathcal{E}(y) = a + bz$ , where  $z = e^{-x}$ . Finally, methods based on linear prediction functions, particularly in this broader context, are justifiable because they often prove to be "nearly as good" as methods based on nonlinear prediction functions.

### 12.3 Linear Regression Functions

If we wish to predict a value of the random variable  $X_0$  from a linear function  $\alpha^* + \beta^*x_1$  of the random variable  $X_1$ , it is reasonable for us to select values  $\alpha^*$  and  $\beta^*$  so as to minimize the mean squared error of prediction, that is, to minimize

$$\mathcal{E}[X_0 - (\alpha^* + \beta^*X_1)]^2, \quad (12.3.1)$$

where expectation is taken over both  $X_0$  and  $X_1$ . By differentiating (12.3.1) with respect to  $\alpha^*$  and  $\beta^*$  (or by algebraic manipulation), it can be shown that (12.3.1) is minimized when

$$\alpha^* = \alpha = \mu_0 - \rho_{01} \frac{\sigma_0}{\sigma_1} \mu_1 \quad \text{and} \quad \beta^* = \beta = \rho_{01} \frac{\sigma_0}{\sigma_1}, \quad (12.3.2)$$

where  $\mu_0 = \mathcal{E}X_0$ ,  $\mu_1 = \mathcal{E}X_1$ ,  $\rho_{01} = \rho(X_0, X_1)$ ,  $\sigma_0 = \sigma(X_0)$ , and  $\sigma_1 = \sigma(X_1)$ . The quantity  $\beta$  is called the *regression coefficient* of  $X_0$  on  $x_1$ . The linear function  $\alpha + \beta x_1$  that minimizes (12.3.1) is called the *linear minimum mean squared error regression function* of  $X_0$  on  $x_1$ , or more simply, the *linear regression function* of  $X_0$  on  $x_1$ . This function is just

$$\mu_0 + \rho_{01} \frac{\sigma_0}{\sigma_1} (x_1 - \mu_1). \quad (12.3.3)$$

Also a simple derivation shows that the minimum value of (12.3.1), which is attained when  $\alpha$  and  $\beta$  are given by (12.3.2), is just

$$\sigma_{0 \cdot 1}^2 \equiv \sigma_0^2(1 - \rho_{01}^2). \quad (12.3.4)$$

Since  $X_0$  and  $\mu_0 + \rho_{01}(\sigma_0/\sigma_1)(x_1 - \mu_1)$  both have mean  $\mu_0$ , the mean error of prediction is zero, and the variance of the error

$$X_0 - \mu_0 - \rho_{01} \frac{\sigma_0}{\sigma_1} (X_1 - \mu_1)$$

is just

$$\mathcal{E}\left[X_0 - \mu_0 - \rho_{01} \frac{\sigma_0}{\sigma_1} (X_1 - \mu_1)\right]^2.$$

Hence  $\sigma_{0 \cdot 1}^2$  is the variance of the errors in predicting  $X_0$  from  $x_1$  using the linear regression function; it is called the *residual variance* of  $X_0$  on  $x_1$ . Since  $\sigma_{0 \cdot 1}^2$  is a strictly decreasing function of  $\rho_{01}^2$ , taking values  $\sigma_0^2$  and zero, respectively, when  $\rho_{01}^2$  is zero and one, it is clear that  $\rho_{01}^2$  is a measure of the linear predictability of  $X_0$  from  $X_1$ .

From (12.3.4), we see that the residual variance  $\sigma_{0 \cdot 1}^2$  of the errors of prediction decreases as the validity coefficient  $|\rho_{01}|$  increases. It must be remembered that this residual variance is an average error variance, or more precisely, the error variance for a randomly selected person. In most situations, however, the conditional error of prediction will vary, often greatly, over the domain of the predictor variable. Therefore, only when it may be assumed

that these conditional errors are approximately equal is it reasonable to consider that this average error of prediction pertains to predictions for prespecified individuals. In typical testing data, the residual variances are smaller in the tails than in the center of the distribution.

The method presented here for the single predictor case may be generalized in a straightforward manner to cover the multipredictor case. Here we wish to predict  $X_0$  from a linear function of the form

$$R(X_0 | x_1, \dots, x_n) \equiv \alpha^* + \beta_1^*x_1 + \beta_2^*x_2 + \dots + \beta_n^*x_n \quad (12.3.5)$$

so as to minimize

$$\mathcal{E}[X_0 - \alpha^* - \beta_1^*X_1 - \beta_2^*X_2 - \dots - \beta_n^*X_n]^2. \quad (12.3.6)$$

The function

$$R(X_0 | x_1, \dots, x_n) = \mu_0 + \sum_{p=1}^n \beta_p(x_p - \mu_p) \quad (12.3.7)$$

that minimizes the mean squared error (12.3.6) is called the *linear regression function* of  $X_0$  on  $x_1, x_2, \dots, x_n$ . We shall give the formulas for  $\beta_p$  in this more general case in Section 12.5.

The (linear) correlation coefficient is the natural measure of association within the theory of linear mean squared error prediction. Solving (12.3.4) for  $\rho_{01}^2$ , we have

$$\rho_{01}^2 = \frac{\sigma_0^2 - \sigma_{01}^2}{\sigma_0^2}, \quad (12.3.8)$$

which makes explicit the interpretation of the correlation coefficient as a measure of association. The correlation coefficient, however, cannot serve as a measure of nonlinear association since, for example, it is possible for two random variables to have a perfect curvilinear relationship and yet have a zero correlation. Also, unless the domain of one random variable coincides, except for a possible linear transformation, with the domain of the second, the two variables cannot have a correlation coefficient of unity, one with the other. In particular, if one of the random variables is continuous and the other discrete, the correlation between them has an upper bound less than unity. We shall discuss this in more detail in Chapter 15.

## 12.4 Multiple and Partial Correlation

The correlation coefficient between the random variable  $X_0$  and its predicted value given by the linear regression function (12.3.7) is called the *multiple correlation coefficient* between  $X_0$  and  $(X_1, X_2, \dots, X_n)$ . Thus the multiple correlation is in fact a standard Pearson product moment correlation between just two random values, namely,  $X_0$  and that linear combination of  $X_1, X_2, \dots, X_n$  which best predicts  $X_0$ . It may also be shown that the multiple correlation

coefficient is the maximum correlation that can be obtained between  $X_0$  and an arbitrary linear function of  $X_1, X_2, \dots, X_n$ . This maximum correlation is attained when the linear function is the linear regression function. This correlation is always nonnegative. In the special case of just two predictor variables, the squared multiple correlation may be expressed as

$$\rho_{0 \cdot 12}^2 = \frac{\rho_{01}^2 + \rho_{02}^2 - 2\rho_{01}\rho_{02}\rho_{12}}{1 - \rho_{12}^2}. \quad (12.4.1)$$

If the true regression of  $X_0$  on  $x_1, x_2, \dots, x_n$  is linear, then the multiple correlation ratio (12.2.3) reduces to the multiple correlation coefficient.

The difference  $E_{0 \cdot 12 \dots n} = X_0 - R(X_0 | x_1, \dots, x_n)$  is called the *residual of  $X_0$  with respect to  $x_1, x_2, \dots, x_n$*  (see 12.3.5). It represents that part of  $X_0$  that remains after the best linear estimate of  $X_0$  in terms of  $x_1, x_2, \dots, x_n$  is subtracted. The residual  $E_{0 \cdot 12 \dots n}$  is uncorrelated with each of the variables  $X_1, X_2, \dots, X_n$ . The variance  $\sigma_{0 \cdot 12 \dots n}^2$  of  $E_{0 \cdot 12 \dots n}$ , the *partial variance of  $X_0$ , given  $x_1, x_2, \dots, x_n$* , is given by (12.5.3). For a single predictor variable, this reduces to (12.3.4).

Let  $E_{0 \cdot 12 \dots (n-1)}$  and  $E_{n \cdot 12 \dots (n-1)}$  be the residuals of  $X_0$  and  $X_n$  with respect to  $X_1, X_2, \dots, X_{n-1}$ . We may regard the correlation between these two residuals as a measure of the correlation between  $X_0$  and  $X_n$  that remains after the removal of the effects of variables  $X_1, X_2, \dots, X_{n-1}$ . This correlation, denoted and defined by

$$\rho_{0 \cdot 12 \dots (n-1)} \equiv \frac{\sigma[E_{0 \cdot 12 \dots (n-1)}, E_{n \cdot 12 \dots (n-1)}]}{\sigma[E_{0 \cdot 12 \dots (n-1)}]\sigma[E_{n \cdot 12 \dots (n-1)}]}, \quad (12.4.2)$$

is called the *partial correlation of  $X_0$  and  $X_n$  with respect to  $X_1, X_2, \dots, X_{n-1}$* . In the simplest case, the squares of the partial correlations are

$$\rho_{01 \cdot 2}^2 = \frac{(\rho_{01} - \rho_{12}\rho_{02})^2}{(1 - \rho_{12}^2)(1 - \rho_{02}^2)}, \quad \rho_{02 \cdot 1}^2 = \frac{(\rho_{02} - \rho_{12}\rho_{01})^2}{(1 - \rho_{12}^2)(1 - \rho_{01}^2)}. \quad (12.4.3)$$

The signs of the partial correlation coefficients are those of the differences in the numerators before they are squared.

The partial regression weights are related to the partial correlations and partial standard deviations in exactly the same way (see 12.3.2) as the corresponding first-order quantities are related, that is, by

$$\beta_{01 \cdot 23 \dots n} = \frac{\sigma_{0 \cdot 12 \dots n}}{\sigma_{1 \cdot 023 \dots n}} \rho_{01 \cdot 2 \dots n}, \quad (12.4.4)$$

where  $\sigma_{i \cdot 01 \dots (i-1)(i+1) \dots n} = \sigma[E_{i \cdot 01 \dots (i-1)(i+1) \dots n}]$ . An equivalent expression is

$$\beta_{01 \cdot 23 \dots n} = \frac{\sigma_{0 \cdot 23 \dots n}}{\sigma_{1 \cdot 23 \dots n}} \rho_{01 \cdot 2 \dots n}. \quad (12.4.5)$$

In the case of two predictor variables  $X_1$  and  $X_2$ , we have

$$\beta_{01 \cdot 2} = \frac{\sigma_0}{\sigma_1} \frac{(\rho_{01} - \rho_{02}\rho_{12})}{(1 - \rho_{12}^2)}, \quad \beta_{02 \cdot 1} = \frac{\sigma_0}{\sigma_2} \frac{(\rho_{02} - \rho_{01}\rho_{12})}{(1 - \rho_{12}^2)}. \quad (12.4.6)$$

The important relation

$$1 - \rho_{0 \cdot 12 \dots n}^2 = (1 - \rho_{01}^2)(1 - \rho_{02 \cdot 1}^2) \cdots (1 - \rho_{0n \cdot 12 \dots (n-1)}^2) \quad (12.4.7)$$

implicitly shows the increment in the multiple correlation as each new predictor variable is added; it shows this increment in terms of the partial correlation that exists between the new predictor and the criterion if the effects of the other predictors are partialled out. This relation also shows that the multiple correlation of any order is at least as large as any of the first-order validities or any of the corresponding partial correlations, that is,  $\rho_{0 \cdot 12 \dots n}^2 \geq \rho_{0n}^2$ ,  $\rho_{0 \cdot 12 \dots n}^2 \geq \rho_{0n \cdot 1}^2$ , and so forth. Equation (12.4.7) also implies the relation

$$1 - \rho_{0 \cdot 12 \dots n}^2 = (1 - \rho_{0 \cdot 12 \dots (n-1)}^2)(1 - \rho_{0n \cdot 12 \dots (n-1)}^2). \quad (12.4.8)$$

## 12.5 Partial and Multiple Correlation and Regression in $n$ Variables\*

When more than two predictor variables are used in regression theory, it becomes convenient to use matrix-algebra notation to obtain concise formulas for the various multiple and partial correlations and the regression weights. Let  $\Sigma = \|\sigma_{pq}\|$  be the covariance matrix of the predictor variables  $X_1, X_2, \dots, X_n$ , and let  $\Sigma^{-1} = \|\sigma^{pq}\|$  be the inverse of  $\Sigma$ , which is assumed to exist. Then it can be shown that the values that minimize (12.3.6) are

$$\alpha^* = \alpha = \mu_0 - \beta_1\mu_1 - \cdots - \beta_n\mu_n, \quad (12.5.1)$$

and

$$\beta_p^* = \beta_p = \sum_{q=1}^n \sigma^{pq}\sigma_{q0}, \quad p = 1, 2, \dots, n, \quad (12.5.2)$$

where  $\mu_i = \mathcal{E}X_i$ ,  $\sigma_{q0}$  is the covariance of  $X_q$  and  $X_0$ , and  $\sigma^{pq}$  is the  $p, q$ th element of  $\Sigma^{-1}$ . The  $p$ th *partial regression weight*  $\beta_p$  is more precisely denoted by  $\beta_{0p \cdot 123 \dots (p-1)(p+1) \dots n}$ .

Let  $\|\sigma_{ij}\|$ ,  $i, j = 0, 1, 2, \dots, n$ , be the covariance matrix of the criterion and the predictors. This matrix differs from  $\|\sigma_{pq}\|$  in that a zeroth row and column giving the variance of the criterion and the covariances between the criterion and the predictors have been added. The minimum mean squared error (12.3.6), denoted by  $\sigma_{0 \cdot 12 \dots n}^2$ , is

$$\sigma_{0 \cdot 12 \dots n}^2 = \frac{|\sigma_{ij}|}{|\sigma_{pq}|}, \quad (12.5.3)$$

where  $|\sigma_{ij}|$  is the determinant of the matrix  $\|\sigma_{ij}\|$ ,  $i, j = 0, 1, 2, \dots, n$  and

---

\* Reading of this section may be omitted without loss of continuity.

where  $|\sigma_{pq}|$  is the determinant of the matrix  $\|\sigma_{pq}\|$ ,  $p, q = 1, 2, \dots, n$ . The quantity  $\sigma_0^2$  is called the (average) residual variance of  $X_0$  on  $x_1, x_2, \dots, x_n$ , or the *partial variance* of  $X_0$ , given  $x_1, x_2, \dots, x_n$ .

We denote the multiple correlation of  $X_0$  with  $X_1, X_2, \dots, X_n$  by  $\rho_{0 \cdot 12\dots n}$ , and we shall find it useful to state several distinct formulas for this quantity. The first of these,

$$\rho_{0 \cdot 12\dots n}^2 = 1 - \frac{|\sigma_{ij}|}{\sigma_0^2 |\sigma_{pq}|}, \quad (12.5.4)$$

expresses the squared multiple correlation in determinantal form, where  $\sigma_0^2$  is the variance of  $X_0$ . A minor variant of (12.5.4) is

$$\rho_{0 \cdot 12\dots n}^2 = 1 - \frac{\sigma_{0 \cdot 12\dots n}^2}{\sigma_0^2}, \quad (12.5.5)$$

which we can obtain from (12.5.3) and (12.5.4).

An alternative form is sometimes more useful. Let  $\mathbf{D}_\sigma^2$  be a diagonal matrix whose diagonal elements are the same as those of  $\Sigma$ . Let  $\mathbf{P}$  be the intercorrelation matrix of predictors; that is,  $\mathbf{P} = \mathbf{D}_\sigma^{-1} \Sigma \mathbf{D}_\sigma^{-1}$ . Let  $\boldsymbol{\rho}$  be the vector of validity coefficients. Then it may be shown that

$$\rho_{0 \cdot 12\dots n}^2 = \boldsymbol{\rho}' \mathbf{P}^{-1} \boldsymbol{\rho}, \quad (12.5.6)$$

or

$$\rho_{0 \cdot 12\dots n}^2 = \boldsymbol{\rho}' \mathbf{D}_\sigma \Sigma^{-1} \mathbf{D}_\sigma \boldsymbol{\rho}. \quad (12.5.7)$$

Consider the validity vector  $\boldsymbol{\rho}' = (0.40, 0.35, 0.05)$ , and the intercorrelation matrix and its inverse,

$$\mathbf{P} = \begin{bmatrix} 1.00 & 0.30 & 0.30 \\ 0.30 & 1.00 & 0.90 \\ 0.30 & 0.90 & 1.00 \end{bmatrix} \quad \text{and} \quad \mathbf{P}^{-1} = \begin{bmatrix} 1.105 & -0.174 & -0.174 \\ -0.174 & 5.291 & -4.709 \\ -0.174 & -4.709 & 5.291 \end{bmatrix}.$$

Then the squared multiple correlation is

$$\rho_{0 \cdot 123}^2 = (0.40, 0.35, 0.05) \begin{bmatrix} 1.105 & -0.174 & -0.174 \\ -0.174 & 5.291 & -4.709 \\ -0.174 & -4.709 & 5.291 \end{bmatrix} \begin{pmatrix} 0.40 \\ 0.35 \\ 0.05 \end{pmatrix} = 0.617.$$

The most convenient and easily remembered expression for the regression weights is

$$\boldsymbol{\beta} = \Sigma^{-1} \boldsymbol{\sigma}, \quad (12.5.8)$$

where  $\boldsymbol{\sigma} \equiv \{\sigma_{0q}\}$  is the vector of covariances of the predictors with the criterion. The squared multiple correlation can then be expressed as

$$\rho_{0 \cdot 12\dots n}^2 = \frac{\boldsymbol{\sigma}' \Sigma^{-1} \boldsymbol{\sigma}}{\sigma_0^2}, \quad (12.5.9)$$

by using (12.5.7), or as

$$\rho_{0 \cdot 12 \dots n}^2 = \frac{\beta' \Sigma \beta}{\sigma_0^2}, \quad (12.5.10)$$

by using (12.5.8) and (12.5.9). Corresponding formulas for the multiple correlations can be obtained by taking positive square roots in (12.5.6), (12.5.7), (12.5.9), and (12.5.10).

The partial correlations can be obtained from (12.4.8) or more directly from

$$\rho_{0n \cdot 12 \dots (n-1)} = -|\rho_{0n}^*|/\sqrt{|\rho_{00}^*| |\rho_{nn}^*|}, \quad (12.5.11)$$

where  $|\rho_{kl}^*|$  is the determinant of the matrix  $\|\rho_{i,j}\|$ ,  $i, j = 0, 1, \dots, n$ , with the  $k$ th row and  $l$ th column omitted. The partial correlations may also be obtained by simple matrix operations. If  $\mathbf{M}$  is a matrix, then let  $(\text{Diag } \mathbf{M})$  be the diagonal matrix whose diagonal elements are identical to those of the matrix  $\mathbf{M}$ . Let

$$\begin{aligned} \mathbf{A} &= \{a_{ij}\} (\text{Diag } \|\rho^{ij}\|)^{-1/2} \|\rho^{ij}\| (\text{Diag } \|\rho^{ij}\|)^{-1/2} \\ &= (\text{Diag } \|\sigma^{ij}\|)^{-1/2} \|\sigma^{ij}\| (\text{Diag } \|\sigma^{ij}\|)^{-1/2}, \end{aligned} \quad (12.5.12)$$

where  $\|\rho^{ij}\|$  and  $\|\sigma^{ij}\|$  are the inverses of the correlation and covariance matrices of the criterion and the predictors. Then

$$a_{ii} = 1 \quad \text{and} \quad a_{ij} = -\rho_{ij \cdot kl \dots}, \quad k, l \dots \neq i \neq j.$$

## 12.6 The Screening of Predictor Variables

It often occurs that when an investigator is faced with a prediction or selection problem he has a very large number of tests available which he can adopt. Although the investigator may have the resources to administer each of these tests for his investigatory study, he usually wants to identify some smaller number of tests for continuing use. Thus his problem is to select  $n$  tests from a larger group of  $N$  tests in such a way that he obtains a satisfactory multiple regression equation. We might state his problem more technically and say that he must select exactly  $n$  tests, choosing them in such a way that no other combination of  $n$  tests provides a higher multiple correlation between the sets of tests and the criterion.

Even if sampling problems are ignored by assuming that all relevant variances and covariances are known, no *simple* algorithm has been found for solving this problem. To obtain (or ensure) the best set of  $n$  variables, the investigator must examine all of the  $N!/n!(N-n)!$  possible choices of tests and compare the multiple correlations obtainable therefrom. This, of course, is a tedious and expensive process and a process that is rarely feasible except for small values of  $N$ , in which case the algorithm given by Garside (1966) is useful.

If the investigator is willing to accept something less than optimality or if he is willing to make some further assumptions about his model, there are computationally less complex techniques available. (We shall discuss these in

Chapter 13.) In many situations, he will find it useful to screen the set of potential predictor variables, either before or instead of adopting one of the formal selection procedures. A study of the formulas for multiple and partial correlation in the three-variable case provides two general rules for such screening procedures.

A simple analysis of the formula (12.4.1),

$$\rho_{0 \cdot 12}^2 = \frac{\rho_{10}^2 + \rho_{20}^2 - 2\rho_{10}\rho_{20}\rho_{12}}{1 - \rho_{12}^2},$$

will indicate how this function varies with its arguments. Differentiating (12.4.1) with respect to  $\rho_{10}$ , we have

$$\frac{\partial \rho_{0 \cdot 12}^2}{\partial \rho_{10}} \propto \rho_{10} - \frac{(\rho_{10} - \rho_{20}\rho_{12})}{1 - \rho_{12}^2} \rho_{20}\rho_{12}.$$

We see that the derivative is positive for  $\rho_{10} > \rho_{20}\rho_{12}$ , and hence that  $\rho_{0 \cdot 12}^2$  is an increasing function of  $\rho_{10}$ . If  $\rho_{10} = \rho_{20}\rho_{12}$ , we have  $\rho_{0 \cdot 12}^2 = \rho_{20}^2$ , in which case variable 1 contributes nothing to the multiple correlation. However, for  $\rho_{10} < \rho_{20}\rho_{12}$ , we see that  $\rho_{0 \cdot 12}^2$  increases as  $\rho_{10}$  decreases. This latter fact proves to be of considerable theoretical and potentially of some applied interest (see Section 12.7). Differentiation with respect to  $\rho_{20}$  yields corresponding results, with the correlations  $\rho_{10}$  and  $\rho_{20}$  interchanged. Differentiation with respect to  $\rho_{12}$  indicates that if

$$\rho_{12} < \min \left\{ \frac{\rho_{10}}{\rho_{20}}, \frac{\rho_{20}}{\rho_{10}} \right\},$$

then  $\rho_{0 \cdot 12}^2$  increases as  $\rho_{12}$  decreases, and if

$$\rho_{12} > \max \left\{ \frac{\rho_{10}}{\rho_{20}}, \frac{\rho_{20}}{\rho_{10}} \right\},$$

then  $\rho_{0 \cdot 12}^2$  increases as  $\rho_{12}$  increases. The reader can gain a more complete understanding of these results by studying Fig. 12.6.1.

Figure 12.6.1 shows the relationship between predictor intercorrelation and multiple correlation in two predictor cases where the higher zero-order correlation is .60 and the lower zero-order correlations are as shown below.

Curve	Correlation of the second predictor with criterion	Curve	Correlation of the second predictor with criterion
1	$\rho_{02} = .10$	5	$\rho_{02} = -.60$
2	$\rho_{02} = .30$	6	$\rho_{02} = -.50$
3	$\rho_{02} = .50$	7	$\rho_{02} = -.30$
4	$\rho_{02} = .60$	8	$\rho_{02} = -.10$

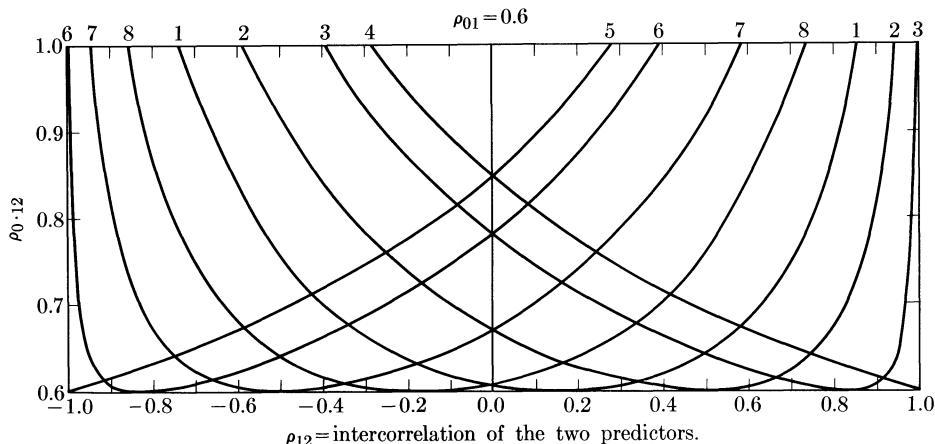


FIG. 12.6.1. Multiple correlation as a function of zero-order coefficients.

The figure illustrates that

- 1) for  $\rho_{12} < \rho_{02}/\rho_{01}$ ,  $\rho_{0.12}$  increases as  $\rho_{12}$  decreases, and for  $\rho_{12} > \rho_{02}/\rho_{01}$ ,  $\rho_{0.12}$  increases as  $\rho_{12}$  increases.
- 2) for  $\rho_{02} > \rho_{01}\rho_{12}$ ,  $\rho_{0.12}$  is an increasing function of  $\rho_{02}$ , and for  $\rho_{02} < \rho_{01}\rho_{12}$ ,  $\rho_{0.12}$  increases as  $\rho_{02}$  decreases. For example: In this figure,  $\rho_{01} = .60$ . Hence, if  $\rho_{02} > .60\rho_{12}$ , then  $\rho_{0.12}$  is an increasing function of  $\rho_{02}$ , and if  $\rho_{02} < .60\rho_{12}$ , then  $\rho_{0.12}$  increases as  $\rho_{02}$  decreases.

This analysis leads to two rules of thumb for the selection of predictor variables.

1. Choose variables that correlate highly with the criterion but that have low intercorrelations.
2. To these variables add other variables that have low or zero correlations with the criterion but that have high correlations with the other predictor variables.

The first rule is of great practical importance. The second is primarily of theoretical interest.

## 12.7 Suppressor Variables, Moderator Variables, and Differential Predictability

Variables that are useful in a regression equation because they have low or zero correlation with the criterion but high correlation with other predictors are called *suppressor variables*. In effect, they represent some aspect of the predictor variables that is not related to the criterion but that functions to

"suppress" or subtract out this invalid component and thus to make the original predictor more valid. The existence of suppressor variables can be seen graphically in Fig. 12.6.1. For example, observe curve 1: If  $\rho_{01} = .60$  and  $\rho_{02} = .10$ , then the multiple correlation is .60 when  $\rho_{12} = \frac{1}{6}$ . This multiple correlation increases as a function of  $\rho_{12}$  when  $\rho_{01}$  and  $\rho_{02}$  are held constant.

Horst (1966) reports encountering a suppressor variable in predicting the success in pilot training during World War II. He reports that tests of mechanical ability, numerical ability, and spatial ability were positively correlated with criterion while a test of verbal ability had a very low positive correlation with criterion. However, the correlations of the verbal ability score with the mechanical, spatial, and numerical ability scores were reasonably high. The partial regression weight for the verbal score was negative, and indeed the inclusion of the verbal score did increase the multiple correlation.

Horst (1966) explains this phenomenon by arguing that verbal ability of a high order was not essential for success in World War II primary pilot training but was necessary to perform successfully on the pencil-and-paper tests of mechanical, spatial, and numerical ability. Thus, Horst argues, "to include the verbal score with a negative weight served to suppress or subtract the irrelevant ability, and to discount the scores of those who did well on the test simply because of their verbal ability". Thorndike (1949), however, reported that in his experience with air-crew selection, those variables that initially appeared to be suppressor variables did not retain this status when larger samples were taken. Despite both the mathematical possibility of the existence of suppressor variables and Horst's report of a suppressor effect in one application, the reader should not expect to "find" such variables regularly, except as a result of sampling fluctuation or as the result of the specific construction of such variables. Dicken (1963), on the basis of his own experimental studies and a review of the published literature, suggested that "... good suppressor variables are hard to find".

It is sometimes said that a suppressor variable is characterized by the fact that it has a negative partial regression weight. This statement requires some qualification. Consider a two-variable regression problem in which the predictor variables  $X_1$  and  $X_2$  have positive regression weights. Let  $Y_1 = -X_1$  and  $Y_2 = X_2$ ; then  $Y_1$  will have a negative and  $Y_2$  a positive partial regression weight. But surely this simple reflection has not changed the underlying structure of the three variables. However, a (pure) suppressor may be defined unambiguously as a variable (1) whose partial regression weight is opposite in sign to its zero-order correlation, when the zero-order correlation is nonzero, or (2) whose partial regression weight is nonzero when its zero-order correlation is zero. It is possible, however, for a variable to act both as a predictor and a suppressor variable simultaneously; its regression weight then has the same sign as its zero-order correlation.

The idea of suppressor variables may have only limited practical significance. The idea of moderator variables, however, seems to promise more in this respect.

For example, as Saunders (1956) has pointed out, it is well known that

- 1) predictability of freshman college grades tends to be higher for women than for men,
- 2) the regression line for predicting college grades of veterans from test scores tends to be different from that for nonveterans, etc.

Knowledge of a student's sex or past military experience would therefore be of value in better predicting his performance in college. Such knowledge, however, may be used in two distinct ways. First, different regression lines may be estimated within each classification defined by these additional variables. Frederiksen and Melville (1954), who have exhibited an instance of the value of this approach, have called it *differential predictability*. Partial replications of this work have been reported by Frederiksen and Gilbert (1960) and by Stricker (1966). One disadvantage of this approach is that the two sets of regression weights must be estimated. It ought to be noted here that the similar-sounding term *differential prediction* is used in a quite different context (see, for example, Horst, 1955, and Mollenkopf, 1950). Differential prediction refers to the ability of a test battery to predict differences in examinee performances on two different criteria.

Second, an assumption may be made about how these additional variables interact with the predictor variables. This approach presumes interactive effects among the predictors and leads to the statement of nonlinear regression equations. It permits the pooling of all data, but of course, it may not be useful if the assumptions of the model are not accurate. An interesting model of this type, suggested by Saunders (1956), yields a nonlinear regression equation involving first-order and product terms but not squared terms: for example,  $\mathcal{E}(z) = ay + bx + cyx$ . Practically, however, we see little justification for not considering the full second-order linear regression function.

Saunders used the term *moderator variables* to describe such variables (Saunders, 1956; Ghiselli, 1963). In differential predictability studies, however, this term has also come to be used to refer to those variables that define distinct groups. A survey of the empirical work using moderator variables and differential predictability is given by Ghiselli (1963). Guion's (1967) review article, "Personnel Selection", is very pertinent to the topic of moderator variables and can be recommended as a good general supplement to this and the next chapter. Lubin and Osburn (1957) have discussed another nonlinear approach, and Cleary (1966) has recently proposed a model that permits a different set of regression weights for each person to emerge.

## 12.8 Incremental Validity

From the discussions in the preceding sections, it should be clear that the utility of a potential predictor variable in a regression equation cannot be determined solely from its zero-order correlation with the criterion. In this

section we shall elaborate this idea somewhat, in terms of the concept of incremental validity. By the *incremental validity* of a variable, we shall mean the degree to which it raises the multiple correlation when it is included in a set of predictor variables. Thus the incremental validity of a variable will differ not only among different criteria but also according to what other predictor variables are chosen. If only one predictor variable is used, then the incremental validity of that variable is simply its usual validity coefficient. This concept of incremental validity is quite natural, and in substance it has been discussed for many years. The adoption of the term, however, seems to have followed from its use in a paper by Sechrest (1963).

Experience in psychology and in many other fields of application has shown that it is seldom worthwhile to include very many predictor variables in a regression equation, for the incremental validity of new variables, after a certain point, is usually very low. This is true because tests tend to overlap in content and consequently the addition of a fifth or sixth test may add little that is new to the battery and still relevant to the criterion. It is not possible to lay down hard and fast rules about how many variables it is profitable to include in the predictor set. That number depends on the relative complexity of the predictors and of the criterion. If the criterion is complex and the predictors are simple, then a larger number of predictors is required than in the opposite situation.

It has been argued (Cronbach and Gleser, 1965, and elsewhere) that the evaluation of the worth of personality tests and interest tests should not be based on a study of their zero-order validities with criteria. These are uniformly low when they are compared to aptitude and achievement tests. The crucial question, it is argued, is: "Do these tests have enough incremental validity when included in a battery of aptitude and achievement tests?" This could be answered affirmatively if these tests reliably measured components of the criterion not measured by the other predictors.

Another point may usefully be made concerning personality tests. Often it is argued that these instruments are useful when employed, in an intuitive-nonstatistical manner, by a skilled clinical practitioner. The consideration of this very reasonable premise, it would seem, merely shifts the subject of the validation study from the test to the practitioner. If this premise is worth considering, then it is worth testing by properly designed experimental studies that would compare

- 1) the accuracy of predictions of clinicians when they do not use tests,
- 2) the accuracy of predictions of clinicians when they do use tests, and
- 3) the accuracy of predictions made from purely statistical treatment of the test results.

The challenge along these lines laid down by Meehl (1955) and sharpened by Sawyer (1966) has not yet been answered satisfactorily. As Meehl (1965) has noted, only one major study (Lindzey, 1965) has shown a clear superiority of clinical over statistical prediction.

### 12.9 Validity and the Selection Ratio

It has long been recognized (see Thurstone, 1932) that the magnitude of a validity coefficient, as such, is an inadequate representation of the value of a test in a selection situation. Another important factor in determining the usefulness of any test is the selection ratio that occurs in the particular selection problem at hand. By the *selection ratio* we mean the ratio of the number of vacancies that must be filled to the number of applicants who have been found acceptable for the position prior to testing, but perhaps after initial screening. This initial screening might consist, say, of the requirement of a high school diploma and the passing of a physical examination.

For present purposes, we shall consider the very special situation where an employee's (or student's) performance is considered to be either satisfactory or unsatisfactory. This can be a very natural and meaningful distinction in some educational situations. In medical schools, for example, the most important immediate criterion is whether or not the student receives his degree (see, for example, Little, Gee, and Novick, 1960). There the utility of a test having a specified validity coefficient depends on the performance that can be expected from a random selection of persons from the prescreened applicant group. If that performance is typically bad, then the test is more useful than if the performance were typically good.

Now it is obvious that if the selection ratio is unity, that is, if there are only as many prescreened applicants as positions, then no test can be useful to us since the assumption is that we shall admit (hire) all applicants regardless of the test results. However, if there are more prescreened applicants than positions, a valid test may prove useful. The rule which linear regression theory requires here is that we select the  $n$  examinees having the highest test scores to fill the  $n$  vacancies. For simplicity, we presume there are no ties. If no test is used, we presume that  $n$  persons are selected randomly from the group of prescreened applicants.

For purposes of illustration only, we now assume a bivariate normal distribution of test scores and performance scores with a specified cutoff on performance scores that separates unsatisfactory from satisfactory performance. Taylor and Russell (1939) have presented a series of tables for this case that show the effects of testing on the proportion of satisfactory performances. Separate tables are presented for cases where the percentage of those considered satisfactory without testing is 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 95. Within each table, the percentage satisfactory with selection based on testing is given as a function of validity and of the selection ratio. Table 12.9.1 is a reproduction of one such table.

This table gives a very striking picture of the potential utility of tests that have a relatively low validity coefficient, if the selection ratio is also low. Note that with a selection ratio of 0.20 and a validity of only 0.20, the use of tests raises the percentage of employees considered satisfactory from 50 to 61. This increase certainly implies a more favorable evaluation of the test than does its

**Table 12.9.1**

Taylor-Russell table of percentage of employees selected by testing  
 considered satisfactory when proportion of employees  
 considered satisfactory before testing is 50%\*

Validity	Selection ratio										
	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0.00	50	50	50	50	50	50	50	50	50	50	50
0.05	54	54	53	52	52	52	51	51	51	50	50
0.10	58	57	56	55	54	53	53	52	51	51	50
0.15	63	61	58	57	56	55	54	53	52	51	51
0.20	67	64	61	59	58	56	55	54	53	52	51
0.25	70	67	64	62	60	58	56	55	54	52	51
0.30	74	71	67	64	62	60	58	56	54	52	51
0.35	78	74	70	66	64	61	59	57	55	53	51
0.40	82	78	73	69	66	63	61	58	56	53	52
0.45	85	81	75	71	68	65	62	59	56	53	52
0.50	88	84	78	74	70	67	63	60	57	54	52
0.55	91	87	81	76	72	69	65	61	58	54	52
0.60	94	90	84	79	75	70	66	62	59	54	52
0.65	96	92	87	82	77	73	68	64	59	55	52
0.70	98	95	90	85	80	75	70	65	60	55	53
0.75	99	97	92	87	82	77	72	66	61	55	53
0.80	100	99	95	90	85	80	73	67	61	55	53
0.85	100	99	97	94	88	82	76	69	62	55	53
0.90	100	100	99	97	92	86	78	70	62	56	53
0.95	100	100	100	99	96	90	81	71	63	56	53
1.00	100	100	100	100	100	100	83	71	63	56	53

\* From H. C. Taylor and J. T. Russell, The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, 1939, **23**, 565-578. Used by permission.

validity coefficient of 0.20. Note also that with a validity of only 0.45 but a selection ratio of 0.05, the use of tests raises the percentage of acceptable employees from 50 to 85.

One must exercise care in using these tables because the situation they describe is highly artificial. The assumption of bivariate normality is certainly inaccurate in most testing applications. The basic situation we have described is highly artificial in another respect as well: We have supposed that we have no valid information on the applicants other than their test scores, and this is seldom the case. For example, in addition to knowing that a student has been graduated from high school, we generally know his class standing and perhaps have a record of prior work experience. It therefore seems that we should use the Taylor-Russell tables to afford perspective, primarily, rather than to supply precise values.

We shall see in Chapter 15 that if a continuous criterion is arbitrarily dichotomized and scored zero-one, then the validity of any predictor varies with the point of dichotomization. Therefore, although the validity coefficient is a useful index of validity, it does have very definite limitations. It is, in fact, only a measure of validity in the sense of linear predictability in the fixed scales of the predictor and the criterion or any linear transformations thereof. For many important problems more specific indexes are required. In the context of and under the assumptions justifying the Taylor-Russell tables, some measure taken from these tables might be more relevant than a validity coefficient. Any such measure, however, would also depend on the dichotomization point. Brogden (1946) and Brown and Ghiselli (1953) have shown the practical utility of validity coefficients in other ways. An integrated discussion of these questions may be found in Cronbach and Gleser (1965).

The conclusion to be drawn from this discussion and the suggested readings is that the validity of a test (in the larger meaning of the word) depends on the particular measure of value that is appropriate to the practical problem at hand.

### **12.10 Some Remarks on the Explication of the Concept of Validity as a Correlation Coefficient**

As we pointed out in the introductory section of this chapter, the explication of the concept of validity as a correlation coefficient is but one possible explication. The development of linear regression and correlation theories in the second, third, and fourth sections now permits us to clarify the very special nature of this particular definition. As we noted in Section 12.4, the multiple correlation between a set of predictors and a criterion (the validity of the predictor battery) is equal to the zero-order correlation between the criterion and the particular linear combination of the predictors that minimizes the squared error of prediction. Thus the multiple correlation is a measure of validity in the specific sense of the minimization of the squared error of prediction.

In effect, multiple regression theory can be considered a special case of statistical decision theory, where the loss function is taken to be proportional to the squared error of prediction. Now there seems little justification for assuming blindly that economic loss in applications is in proportion to the squared error of prediction, and the adoption of the squared error loss function might therefore seem unjustified. However, the adoption of squared error loss leads to using a mean value as an estimator. And it is true, as pointed out by Chernoff and Moses (1959, pp. 207 ff.), for example, that this mean value estimator is either optimal or nearly optimal for a large class of loss functions and a large class of distributions. Certainly the mean value estimator is a reasonably robust estimator. Moreover we may recall that in Chapter 3 we pointed out that the length of the Chebychev confidence bound (or the confidence bound based on a normal error distribution) is directly proportional to

the standard error of prediction. Thus we see that linear (or nonlinear) regression and correlation theory is typically a reasonable explication of the concept of validity when our interest is in minimizing the uncertainty of our prediction of performance scores over all examinees.

In many applications, however, we have little interest in predicting the exact score that a person will attain on a criterion. Our only concern is in predicting whether or not his performance on the criterion will be above or below some critical level. For example, in applying the Taylor-Russell tables we are concerned only with whether or not an examinee's future performance is above some critical value that distinguishes those who are acceptable from those who are unacceptable. We found that in this case the validity (in the general sense) of a test is not directly proportional to its correlation with the criterion although, for a fixed selection ratio, it is monotonically related to this correlation.

This problem actually introduces a new loss function, one according to which there is no loss if a selected examinee is successful and there is a unit loss if he is not. This same kind of loss function underlies the work in Chapter 19 which deals with classifying examinees in high or low ability groups. When the methods of Chapters 16 through 20 are used to construct a test, we shall see that it is not meaningful to evaluate the excellence of the resulting test on the basis of its validity coefficient. In fact these methods, when successful, can produce tests with very low validity coefficients. In such situations other measures of validity are required.

## 12.11 Construct Validity

For scientific purposes, the most important characteristic of a test is its *construct validity*. By construct validity we mean the degree to which a test measures the construct it was designed to measure. The subject of construct validity is a difficult and controversial one (Cronbach and Meehl, 1955; Loevinger, 1957; Bechtoldt, 1959; Campbell and Fiske, 1959), which has been only partially explicated both conceptually and mathematically.

The difficulty in establishing the construct validity of a test is that the criterion, the construct, is not directly measurable. Hence a test-criterion correlation cannot be computed. However, it is possible to compute correlations of the test with other tests, some of which, according to theory, should correlate with the construct and some of which should not. If two tests correlate with each other, there is the suggestion that it is because they both, in part, measure the construct. Thus construct validity may be studied, though only indirectly, by analyzing observed-score correlations.

We shall concern ourselves here and in Section 15.10 entirely with correlational methods useful in studying construct validity. Here we shall discuss the construct validation of a given test; in Section 15.10 we shall consider the question of construct validity in the context of test construction.

Two important steps are required to establish the construct validity of a test. First it is necessary to show that the test correlates appreciably with all other tests with which theory suggests it should correlate. Then it is necessary to show that the test does not correlate appreciably (except perhaps "spuriously") with all other tests with which theory suggests it should not correlate.

In carrying out these two steps, we must recognize that two tests may correlate "spuriously" because they both measure, in part, something other than the construct of interest. An area of major concern in the construct validation of personality scales is that of response styles. Often, for example, two personality scales correlate only because of "yea-saying" behavior on the items of each scale. To control for this "extraneous" correlation it is necessary to show that the theoretically suggested correlations remain even after the effects of this response style variable have been partialled out. Conversely, to show that two tests measure different constructs it may be necessary to partial out the effects of response style variables that cause a spurious correlation between them. Some considerations in implementing construct validity studies are discussed by Campbell and Fiske (1959), who have suggested the terms *convergent* and *discriminant validity* to describe the two components of construct validity. The subject of construct validity is likely to continue to undergo extensive investigation.

Although the regression methods discussed in this chapter focus in each case on a *single* criterion, it is clear that *many* criteria are of interest in most practical situations, as we have indicated in the Introduction. Unfortunately no entirely satisfactory statistical method of handling multiple criteria is available. However, two important contributions have been made by Horst (1954, 1955). If an investigator chooses to study a single criterion, it is often simply a necessary expediency. Usually only secondarily important criteria such as grade point average and performance on standardized achievement tests are readily available for statistical analysis. More meaningful criteria, such as *lifetime contribution to a profession*, are typically unavailable until long after any benefit might accrue from the study.

### Exercises

- 12.1. Using the notation of Section 12.5 and assuming that  $X_0, X_1, \dots, X_n$  each have unit variance, show that  $\rho_{0 \cdot 12 \dots n}^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$ .
- 12.2. Show that in the special case of three predictor variables, the squared multiple correlation can be expressed as

$$\begin{aligned}\rho_{0 \cdot 123}^2 &= [\rho_{01}^2(1 - \rho_{23}^2) + \rho_{02}^2(1 - \rho_{13}^2) + \rho_{03}^2(1 - \rho_{12}^2) + 2\rho_{01}\rho_{02}(\rho_{13}\rho_{23} - \rho_{12}) \\ &\quad + 2\rho_{01}\rho_{03}(\rho_{12}\rho_{23} - \rho_{13}) + 2\rho_{02}\rho_{03}(\rho_{12}\rho_{13} - \rho_{23})] \\ &\quad \times [1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}]^{-1}.\end{aligned}$$

- 12.3. Show that in the special case of three predictor variables, the squared second-order partial correlation can be expressed as

$$\rho_{03 \cdot 12}^2 = \frac{(\rho_{03 \cdot 1} - \rho_{02 \cdot 1}\rho_{32 \cdot 1})^2}{(1 - \rho_{02 \cdot 1}^2)(1 - \rho_{32 \cdot 1}^2)},$$

where  $\rho_{03 \cdot 1}$ ,  $\rho_{02 \cdot 1}$ , and  $\rho_{32 \cdot 1}$  are first-order partial correlations.

- 12.4. Using the result of the previous exercise and Eq. (12.4.3), show that if the appropriate first-order partial correlations are substituted for the zero-order coefficients, then the second-order partial correlations are obtained. Conclude that in general the squared  $(n - 1)$ -order partial correlation coefficient can therefore be expressed as

$$\rho_{0n \cdot 123 \dots (n-1)}^2 = \frac{[\rho_{0n \cdot 12 \dots (n-2)} - \rho_{0(n-1) \cdot 12 \dots (n-2)}\rho_{n(n-1) \cdot 12 \dots (n-2)}]^2}{[1 - \rho_{0(n-1) \cdot 12 \dots (n-2)}^2][1 - \rho_{n(n-1) \cdot 12 \dots (n-2)}^2]}.$$

- 12.5. If  $\rho_{01} = .60$  and  $\rho_{02} = .20$ , how high must  $\rho_{12}$  be if variable 2 is to act as a suppressor variable in the multiple correlation  $\rho_{0 \cdot 12}$  and raise this multiple correlation to .65?

- 12.6. Suppose you are given the intercorrelation matrix of a set of predictor variables and their reliabilities and validities with respect to a fixed criterion. Derive a formula for the best lower bound on the reliability of the criterion.

- 12.7. It is frequently of interest to determine whether or not a square symmetric matrix of values  $0 \leq \rho_{ij} \leq 1$ ,  $\rho_{ii} = 1$  could be a correlation matrix. A necessary and sufficient condition for this to be true is that the matrix be positive definite. This implies that all leading principal minors are positive. Let  $\mathbf{A} = \{a_{ij}\}$  be a matrix. Then the leading principal minors are

$$p_1 = a_{11}, \quad p_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad p_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \dots, \quad p_n = |a_{ij}|.$$

Suppose that we are given the matrix

$$\mathbf{A} = \begin{vmatrix} 1 & 0.93 & 0.91 \\ 0.93 & 1 & a_{23} \\ 0.91 & a_{32} & 1 \end{vmatrix},$$

with unknown element  $a_{23}$ . Show that if  $\mathbf{A}$  is to be a correlation matrix, then  $a_{23} \geq 0.6926$  (approximately).

- 12.8. Find the regression function of  $T_1$  on  $X_1$ ,  $X_2$ . Find its residual variance.

## References and Selected Readings

ANDERSON, T. W., *An introduction to multivariate statistical analysis*. New York: Wiley, 1948.

BEATON, A. E., The use of special matrix operators in statistical calculus. *Research Bulletin 64-51*. Princeton, N.J.: Educational Testing Service, 1964.

- BECHTOLDT, P., Construct validity: a critique. *American Psychologist*, 1959, **14**, 619-629.
- BROGDEN, H. E., On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, 1946, **37**, 65-76.
- BROWN, C. W., and E. E. GHISELLI, Per cent increase in proficiency resulting from use of selective devices. *Journal of Applied Psychology*, 1953, **37**, 341-344.
- CAMPBELL, D. T., and D. W. FISKE, Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81-105.
- CHERNOFF, H., and L. E. MOSES, *Elementary decision theory*. New York: Wiley, 1959.
- CLEARY, T. ANNE, An individual differences model for multiple regression. *Psychometrika*, 1966, **31**, 215-224.
- COCHRAN, W. G., Improvement by means of selection. In J. Neyman (Ed.), *Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1950.
- CRAMÉR, H., *Mathematical methods of statistics*. Princeton, N.J.: Princeton University Press, 1946.
- CRONBACH, L. J., Response sets and test validity. *Educational and Psychological Measurement*, 1946, **6**, 475-494.
- CRONBACH, L. J., and GOLDINE C. GLESER, *Psychological tests and personnel decisions*. Urbana, Ill.: University of Illinois Press, 1965.
- CRONBACH, L. J., and P. E. MEEHL, Construct validity in psychological tests. *Psychological Bulletin*, 1955, **52**, 281-302.
- DICKEN, C., Good impression, social desirability, and acquiescence as suppressor variables. *Educational and Psychological Measurement*, 1963, **23**, 699-720.
- EDWARDS, A. L., *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- FREDERIKSEN, N., and A. C. F. GILBERT, Replication of a study of differential predictability. *Educational and Psychological Measurement*, 1960, **20**, 759-767.
- FREDERIKSEN, N., and S. D. MELVILLE, Differential predictability in the use of test scores. *Educational and Psychological Measurement*, 1954, **14**, 647-656.
- FREEMAN, H., *Introduction to statistical inference*. Reading, Mass.: Addison-Wesley, 1963.
- GARSHIDE, M. J., The best sub-set for multiple regression analysis. *Applied Statistics*, 1966, **14**, 196-200.
- GHISELLI, E. E., Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educational and Psychological Measurement*, 1960, **20**, 675-684.
- GHISELLI, E. E., Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, 1963, **47**, 81-86.
- GUION, R. M., Personnel selection. *Annual Review of Psychology*, 1967, **18**, 191-216. Palo Alto: Annual Reviews, Inc.
- HOHN, F. E., *Elementary matrix algebra*, 2nd Edition. New York: Macmillan, 1965.

- HORST, P., A technique for the development of a differential prediction battery. *Psychological Monograph*, 1954, No. 380.
- HORST, P., A technique for the development of a multiple absolute prediction battery. *Psychological Monograph*, 1955, No. 390.
- HORST, P., *Psychological measurement and prediction*. Belmont, Calif.: Wadsworth, 1966.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. Vol. 2: *Inference and relationship*. London: Griffin, 1961.
- LINDZEY, G., Seer versus sign. *Journal of Experimental Research in Personality*, 1965, **1**, 17-26.
- LITTLE, J. M., HELEN H. GEE, and M. R. NOVICK, A study of the Medical College Admissions Test in relation to academic difficulties in medical school. *Journal of Medical Education*, 1960, **35**, 264-272.
- LOEVINGER, JANE, Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, **3**, 635-694 (Monograph Supplement No. 9).
- LUBIN, A., Some formulae for use with suppressor variables. *Educational and Psychological Measurement*, 1957, **17**, 286-296.
- LUBIN, A., and H. G. OSBURN, A theory of pattern analysis for the prediction of a quantitative criterion. *Psychometrika*, 1957, **22**, 63-73.
- MEEHL, P. E., *Clinical vs. statistical prediction*. Minneapolis: University of Minnesota Press, 1955.
- MEEHL, P. E., Seer over sign: the first good example. *Journal of Experimental Research in Personality*, 1965, **1**, 27-32.
- MEEHL, P. E., and A. ROSEN, Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 1955, **52**, 194-216.
- MOLLENKOPF, W. G., Predicted differences and differences between predictions. *Psychometrika*, 1950, **15**, 409-417.
- OLKIN, I., and J. W. PRATT, Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 1958, **29**, 201-211.
- OSBURN, H. G., and A. LUBIN, The use of configural analysis for the evaluation of test scoring methods. *Psychometrika*, 1957, **22**, 359-371.
- PETERS, C., and W. VAN VOORHIS, *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
- RIDER, P. R., *An introduction to modern statistical analysis*. New York: Wiley, 1939, pp. 126-128.
- ROZEBOOM, W. W., *Foundations of the theory of prediction*. Homewood, Ill.: The Dorsey Press, 1966.
- SAUNDERS, D. R., Moderator variables in prediction. *Educational and Psychological Measurement*, 1956, **16**, 209-222.
- SAWYER, J., Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 1966, **66**, 178-200.
- SECHREST, L., Incremental validity: a recommendation. *Educational and Psychological Measurement*, 1963, **23**, 153-158.

- STRICKER, L. J., Compulsivity as a moderator variable: a replication and extension. *Journal of Applied Psychology*, 1966, **50**, 331-335.
- TAYLOR, H. C., and J. T. RUSSELL, The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, 1939, **23**, 565-578.
- THORNDIKE, R. L., *Personnel selection*. New York: Wiley, 1949.
- THURSTONE, L. L., *The reliability and validity of tests*. Ann Arbor, Mich.: Edwards Brothers, 1932.
- WHERRY, R. J., Test selection and suppressor variables. *Psychometrika*, 1946, **4**, 239-248.
- WILKS, S. S., *Mathematical statistics*. New York: Wiley, 1963 (second printing with corrections).

## CHAPTER 13

# THE SELECTION OF PREDICTOR VARIABLES

### 13.1 Introduction

The typical approach to the problem of the prediction of job or academic performance employs multiple regression techniques. As we indicated in the previous chapter, standard nonlinear regression methods and some other prediction methods, such as those using moderator variables, can be reduced to problems in linear multiple regression.

In Chapters 6 and 12, we discussed some of the problems involved in comparing two competing tests. In this chapter, we shall be concerned with a slightly different problem. We assume that we are able initially to obtain some large number of potential predictors and wish to select some smaller number for use on a continuing basis. After discussing some relevant sampling problems in the next section, we describe two convenient procedures for selecting some smaller set of predictor variables from a larger set. In Section 13.4, we shall briefly discuss the very difficult problem of using data from a small sample to make predictions about a second sample, and we shall illustrate many of the techniques discussed in this and the previous chapter with a typical validity study.

In Section 13.5, we shall derive formulas for the effect of changes in test length on reliability, validity, and covariance matrices of predictor variables for the multipredictor case. These results are then used in conjunction with a very general problem in predictor variable selection. For this problem, we assume that the length (in time or number of items, as appropriate) of each predictor may be increased or decreased as required. We then determine the amount of time that, for optimal prediction, should be assigned to each predictor under the restriction that the total testing time is some fixed constant.

### 13.2 Some Sampling Problems

The development in the previous chapter involved population parameters, which we assumed to be known exactly. In practice, of course, this is never the case and often the sample data at hand are not of enormous size. This introduces further problems, some of which have not yet proved amenable to analytic solution. In this and the next section, we shall discuss some of these problems,

the few analytic results which bear on them, and the inference procedures that have been adopted to partially overcome them.

If variables are selected for a prediction battery on the basis of sample correlations, the investigator will be selecting not only on the basis of true correlation but also on the basis of sampling errors. In Section 3.7, we showed that the regression model yields an expected true score that is less than the examinee's observed score for all observed scores greater than the mean population observed score. Similarly it is apparent that the regression estimate of true correlation is lower than the obtained sample correlation for observed correlations that are greater than the average of the observed correlation. Since selection typically involves choosing some small number of variables with the highest sample correlations, the effect of selection is usually to overestimate the true multiple correlation for the variables selected. We describe this by saying that there is a "capitalization on chance" when variables are selected in this way. Typically a "shrinkage" in the multiple correlation is generally found when these variables are used on a new sample.

Because the basic statistical problems in this area remain unsolved, an empirical approach to the problem is necessary. One commonly used procedure is to select variables on the basis of one sample, the *screening sample*, and then to estimate the multiple correlation and regression weights for these selected variables on the basis of a second sample, the *calibration sample*, for which the predictor variables have been prespecified. If variables are selected and regression weights estimated in a single sample, then it generally is advisable to apply these variables and weights in a new sample to see how valuable the specified composite is. This is called *cross validation*.

Even with a moderate sample size, the amount of shrinkage in the multiple correlation between the screening and calibration or cross validation samples may, in fact, be substantial if a very small number of predictor variables has been selected from a very large set of potential predictors. If, additionally, the sample size in the screening sample is small, the shrinkage can be nearly total.

If the reader has been left unworried by the above remarks, he should acquaint himself with the shrinkage encountered by Mosteller and Wallace (1964, Chapter 5) in their discriminant function analysis of the Federalist papers. Had they not had the foresight to reserve some of their data for use as a calibration sample, they would have overestimated the true "discrimination index" for their discriminators *by more than 50%*. Also, two cross validation studies are described in Section 13.5; in one case the shrinkage is far less drastic, in the other the shrinkage is almost total.

Actually, because of errors due to sampling of persons, the sample squared multiple correlation coefficient for *prespecified* variables has a positive bias as an estimator of the true multiple correlation. When the random variables have a multivariate normal distribution, however, a correction for *this* bias is possible. Olkin and Pratt (1958) have derived an unbiased estimator of the squared

multiple correlation. For most practical use, the approximation

$$\widehat{\rho^2} = r^2 - \frac{n-2}{N-n-1} (1-r^2) - \frac{2(N-3)}{(N-n-1)(N-n+1)} (1-r^2)^2, \quad (13.2.1)$$

where  $r^2 = r_{0\cdot 12\dots n}^2$  is the sample multiple correlation,  $n$  is the number of predictors, and  $N$  is the sample size, will be satisfactory. An unfortunate feature of the unbiased estimator and the approximation to it (and many other unbiased estimators of positive quantities) is that they may sometimes take on negative values.

In contrast to the correlation coefficient, very simple unbiased estimates of the regression weights are available. If the theory of minimum mean squared error is applied in a sample, the resulting estimates (which in this case are just the sample regression weights) provide unbiased estimates of the corresponding population quantities. These estimates are called *least squares estimates*. The application of the method of least squares to the estimation of variances and partial variances leads to the unbiased estimate  $[N/(N-n-1)]s^2$ , where  $s^2$  is the relevant sample partial variance,  $N$  is the sample size, and  $n$  is the number of predictor variables. For the zero-order variance ( $n=0$ ), this reduces to the familiar form  $[N/(N-1)]s^2$ .

Wherry (1940) has also provided a simple, alternative (though less accurate) approximate correction for the bias in the sample correlation coefficient; it also provides a reasonable working rule for deciding whether or not to include an additional variable or variables in a regression equation. The squared multiple correlation may be written as (12.5.5):

$$\rho_{0\cdot 12\dots n}^2 = 1 - \frac{\sigma_{0\cdot 12\dots n}^2}{\sigma_0^2}. \quad (13.2.2)$$

If we denote the corresponding sample variance and partial variance by  $s_0^2$  and  $s_{0\cdot 12\dots n}^2$ , respectively, the sample multiple correlation will be given by

$$r_{0\cdot 12\dots n}^2 = 1 - \frac{s_{0\cdot 12\dots n}^2}{s_0^2}. \quad (13.2.3)$$

Now, replacing the numerator and denominator of the second term on the right-hand side of (13.2.2) by their unbiased estimates  $[N/(N-n-1)]s_{0\cdot 12\dots n}^2$  and  $[N/(N-1)]s_0^2$ , we obtain the estimate

$$\begin{aligned} \widehat{\rho^2} &= 1 - \frac{\left(\frac{N}{N-n-1}\right)s_{0\cdot 12\dots n}^2}{\left(\frac{N}{N-1}\right)s_0^2} = 1 - \left(\frac{N-1}{N-n-1}\right)(1-r_{0\cdot 12\dots n}^2) \\ &= \frac{(N-1)r_{0\cdot 12\dots n}^2 - n}{(N-n-1)}, \end{aligned} \quad (13.2.4)$$

where  $N$  is the sample size,  $n$  is the number of predictor variables, and  $N > n + 1$ . This differs slightly from Wherry's original formula, which has the value  $N$  where  $N - 1$  is found in the numerator and denominator of (13.2.4).

Formula (13.2.4) and the form originally given by Wherry have been called Wherry's *correction for shrinkage*. This terminology is confusing and undesirable because this "correction" has nothing at all to do with the capitalization on chance that occurs when a sample multiple correlation is obtained from *selected* variables, nor with the resulting shrinkage in the multiple correlation obtained in the calibration sample.

This correction can be justified from a different point of view. Under an assumption of multivariate normality, or otherwise asymptotically under a very broad assumption, it is true that for  $n > i$ ,

$$\frac{(r_n^2 - r_i^2)/(n - i)}{(1 - r_n^2)/(N - n - 1)} \quad (13.2.5)$$

is distributed as  $F$  with  $n - i$  and  $N - n - 1$  degrees of freedom,  $r_n^2$  and  $r_i^2$  being the sample squared multiple correlations based respectively on  $n$  and on any prespecified  $i$  of the given  $n$  variables (Kendall and Stuart, 1961). Now suppose  $F = 1$ . Then

$$\frac{(r_n^2 - r_i^2)}{n - i} = \frac{(1 - r_n^2)}{N - n - 1}. \quad (13.2.6)$$

Suppose then that the Wherry correction formula is applied to  $r_n^2$  and  $r_i^2$ , and that the two resulting corrected squared multiple correlations are equal; that is,

$$\frac{(N - 1)r_n^2 - n}{N - n - 1} = \frac{(N - 1)r_i^2 - i}{N - i - 1}. \quad (13.2.7)$$

Then we can easily see that (13.2.6) and (13.2.7) are equivalent (Exercise 13.3). Now consider a procedure that adds predictors one at a time in an order that maximizes the incremental validity at each step. The procedure stops adding variables when the corrected squared multiple for the larger set is less than the corrected squared multiple for the smaller set. Compare this with a procedure based on the rule to stop adding variables when the variance ratio (13.2.5) is less than one. Clearly the two procedures are equivalent. Thus the Wherry correction has a reasonable theoretical justification although this justification is not associated with the concept of shrinkage resulting from the selection of variables.

It should also be pointed out that  $\sigma_{0 \cdot 12 \dots n}^2$  is the variance of the errors made in the population when predicting  $X_0$  from the known best linear combination of  $x_1, x_2, \dots, x_n$ . In practice, however, the true regression weights are unknown, and instead a set of least squares estimates of these regression weights based on a prior sample must be used to define a linear prediction function. These estimated regression weights will almost never be the true regression

weights; hence, when this linear prediction function is used, the variance of the errors of prediction in the population will almost always be greater than  $\sigma_{0 \cdot 1 \cdot 2 \dots n}^2$ , the variance of the errors of prediction when the linear regression function is used. Thus the usual estimate of the partial variance is not an unbiased estimate of the error variance associated with the use of this linear prediction function, but only an unbiased estimate of the error variance associated with the use of the true, but unknown, linear regression function. On the average, when the estimated regression function is used in a new sample, the error variance will be greater than the estimated residual variance for the reasons already given. Because of sampling variation, however, the residual variance in any particular second sample may in fact take the extreme value zero, on the one hand, or a value equal to the variance of the criterion, on the other hand.

### 13.3 Formal Procedures for Selecting Predictor Variables

Even if sampling problems could be ignored, the only way to be sure of obtaining the best  $n$  of  $N$  predictors would be to determine the multiple correlation for every such set. This *exhaustion procedure* can seldom be justified economically unless  $N$  is very small. There are two basic formal algorithms for selecting a "good" set of  $n$  predictor variables from a larger set of  $N$  possible predictor variables. The first of these, associated with the names of Wherry (1940), Dwyer (1945), and Summerfield and Lubin (1951), may be called the *forward selection procedure*. It involves a sequential selection of predictor variables such that the predictor variable selected at each stage is the one that provides the largest incremental validity, given all the predictor variables previously selected. In the first stage, this results in the selection of the variable having the highest zero-order correlation with the criterion. In the second stage, the variable selected is the one that has the largest partial correlation with the criterion when the first selected predictor is partialed out. This pair of variables gives the largest multiple correlation among all pairs of variables that include the variable selected in the first stage. However, it is possible to show that this pair of variables does not necessarily provide the highest multiple correlation over all possible pairs of predictor variables. We may show that in general the addition of the  $(n + 1)$ -variable with the highest incremental validity does not necessarily yield the best set of  $(n + 1)$  predictor variables. So far as the present writers have been able to determine, no analytic results have ever been provided to show just how efficient the forward method is for typical problems.

Summerfield and Lubin (1951) have presented what appears to be the most reasonable approach to the problem of deciding when to stop adding variables to the predictor set. In their method, the  $F$ -statistic (13.2.5) is computed at each stage, with  $n = n$  and  $i = n - 1$ , to determine whether or not the additional variable is indeed contributing to prediction. A second  $F$ -statistic is also computed, with  $n$  equal to the total number of variables in the pool and  $i$  equal to the number of variables so far selected, to determine whether or not

the remaining variables, in combination, contribute to prediction. The evaluation of this second  $F$ -ratio is designed to discover any errors in the forward method, errors possibly due to the existence of suppressor variables.

A second procedure, the *backward elimination procedure* described by Horst and MacEwan (1960), begins with all  $N$  variables and then successively eliminates variables so that the decrease in the multiple correlation is minimized at each stage. The problem encountered with the forward method is also to be found in this procedure, for we have no guarantee that we indeed have the best combination of predictors at any stage past the first. Unless the number of variables to be selected is very near the total number in the predictor pool, the forward procedure involves less computation; and certainly the computation will be very much less if only a small percentage of predictors from the pool is to be retained. Since all computational methods in effect involve an inversion of the matrix of predictors, problems of ill-conditioning of this matrix (the matrix being too nearly singular for computational purposes) are more likely to occur with more rather than with fewer variables. If this happens, then the backward procedure cannot be used, but the forward procedure "will provide usable regression equations prior to degeneracy" (Efroymson, 1966).

For the data given in Section 12.5,  $X_1$  is the best single predictor. If  $X_2$  is taken in combination with  $X_1$ , a higher multiple correlation is obtained than if  $X_3$  is taken. The multiple correlations are 0.467 and 0.405, respectively. Hence the forward selection procedure takes  $X_1$  and  $X_2$  as the best two-variable set. However, the optimal two-variable set is  $X_2$  and  $X_3$ , for which the multiple correlation is 0.702, and this set would be the one selected by the backward procedure.

A refinement of the forward and backward procedures called *a stepwise procedure* has proved useful. For the forward procedure, briefly, this refinement is based on reevaluation of each member of the set of selected predictors every time a new predictor is added to the set. "A variable which may have been the best single variable to enter at an early stage may, at a later stage, be superfluous because of the relationships between it and other variables now in the regression" (Draper and Smith, 1966). If this is the case, this variable may be eliminated from the regression equation. A similar refinement is applicable to the backward procedure. Swoyer (1966) has used the backward stepwise procedure and obtained some encouraging results.

### 13.4 Prediction in Future Samples

In practice, regression weights are never known exactly; they must be estimated from a calibration sample. These estimates are then substituted for the true but unknown regression weights in the linear prediction model. The resulting linear prediction equation is then used to "predict" values of the criterion for other individuals, given measurements on the predictors. Even if this calibration sample is distinct from an initial screening sample, not all problems are

solved. There will still be some tendency to capitalize on chance in estimating the regression weights for these variables.

If the true regression weights were known, an increase in the number of predictor variables could never result in a decrease in precision of prediction. However, if the true regression must be estimated, and particularly if the calibration sample is not substantially larger than the number of predictor variables, it can happen that an increase in the number of variables results in a decrease in the precision of prediction for individuals not in the calibration sample. Thus we may have

$$\mathcal{E} \left( Y - \sum_{i=1}^M \hat{\beta}_i X_i \right)^2 \quad \text{greater than or less than} \quad \mathcal{E} \left( Y - \sum_{i=1}^{M+N} \hat{\beta}_i X_i \right)^2,$$

depending (1) on the incremental validity of the last  $N$  predictor variables and (2) on the loss of precision of estimation due to the introduction of  $N$  additional parameters to be estimated. In the extreme, if a linear prediction function that has been determined from a very small calibration sample is used for prediction in a new sample, then it can happen that the expected variance error of prediction is larger than the variance of the criterion. In such cases, an investigator would do better to discard his predictors and use the sample mean value of the criterion as his predicted value.

The predictor-variable selection procedures described in the preceding section are often used as methods for deciding on the specification of variables to be included in the predictor set. Other approaches, advanced by Burket (1964), Elfving (1961), Elfving, Sitgreaves, and Solomon (1961), and Horst (1941), are based on the assumption of an underlying "factor" structure (see Chapter 24), which in effect involves a reduction in the rank of the prediction system (see Section 16.7). These methods seem promising for use in large scale studies. Very recent work of Fortier (1966a, b) should also be studied. Another interesting approach to this problem is that of Linhart (1960), who assumes a normal distribution of errors and then specifies a stopping rule for the forward selection procedure based on the criterion of the minimization of the confidence interval for a future observation. Papers of Stein (1960) and Nicholson (1960) are also pertinent. At present, however, no entirely satisfactory solution to this problem is available.

Rydberg (1963) and others have suggested that another problem may arise when regression weights obtained in one sample are used for prediction in a new sample. Often the determination of the beta weights is made on a group preselected on the basis of some of the potential predictor variables. If no allowance is made for such preselection, then variables so used in selection will typically have drastically reduced beta weights with the reduction being greatest for the best variables. The application of such weights to an unselected group could produce unsatisfactory results.

We may illustrate many of the techniques described in this and the previous chapter with some data from a simple yet effective validity study from the

**Table 13.4.1**  
Data from the 1959–60 independent school  
SSAT prediction study

	V	1960 Correlations					Mean		Standard deviation	
		Q	T	PAS	GPA	1959	1960	1959	1960	1959
Correlations	V	—	.2339	.8681	.2493	.3859	294.0	294.4	13.0	12.6
	Q	.3231	—	.6670	.3059	.4019	318.1	313.8	13.6	11.6
	T	.8771	.7188	—	.3519	.5088	304.0	302.4	10.3	9.2
	PAS	.2212	.2564	.2706	—	.4502	88.2	88.1	5.7	6.2
	GPA	.3901	.4753	.5223	.4271	—	74.5	75.5	7.1	7.4

statistical report *Secondary School Admission Test (SSAT) Scores as Predictors of Ninth Grade Averages, 1959–60 and 1960–61* by Barbara Pitcher of Educational Testing Service. We shall discuss only a small portion of that study here.

A sample of 109 ninth-grade enrollees was obtained in an independent secondary school in 1959 and a second sample of 120 from the same school in 1960. The previous average school marks (PAS) and the Verbal (V), Quantitative (Q), and Total (T) scores,  $T = V + Q$ , were among the predictors available for each enrollee. The tests had been administered in the previous year. Although several performance criteria were available, we shall consider only the overall end-of-year ninth-grade grade point averages (GPA). The data that were obtained in the first and second samples have been summarized in Table 13.4.1.

It should be observed that in most instances the 1960 correlations (the above-diagonal entries) are very close to the corresponding 1959 correlations (the below-diagonal entries). Also it should be noted that the 1959 and 1960 means and standard deviations are remarkably close; this indicates that there was little difference in the quality of the two entering classes.

Test-score validities found in this school follow a pattern typically found in validity studies of this kind. These validities are quite satisfactorily high, considering that they were obtained from the selected rather than the applicant group (see Sections 6.8 through 6.10).

Multiple correlations and regression weights were computed for several combinations of predictors. These values are given in Table 13.4.2. The combination of previous average either with GSAT-T or with GSAT-V and GSAT-Q provides a cross-validated multiple correlation of 0.60 *in the selected group*.

The final column of this table shows a particularly interesting feature of the analysis. The regression weights obtained from the 1959 (1960) sample were used to predict criterion scores from the corresponding predictors in the 1960 (1959) sample, the computations being carried out according to (4.7.3). These computations yielded the cross-validated multiple correlations in the

**Table 13.4.2**

Regression weights, multiple correlations, and  
cross-validated multiple correlations for several  
combinations of predictor variables

Year	GSAT T	GSAT V	GSAT Q	Previous average	Multiple correlation	Cross-validated composite correlation
1959	0.3013			0.3804	0.6008	0.5852
1960	0.3220			0.3717	0.5855	0.6005
1959		0.1431	0.2021		0.5370	0.5001
1960		0.1818	0.2103		0.5017	0.5357
1959		0.1184	0.1706	0.3631	0.6062	0.5785
1960		0.1473	0.1582	0.3747	0.5802	0.6044

final column. It should be noted that any shrinkage found here arises only from variations of weights and not from selection of variables during the study; on the contrary, the variables were chosen ahead of time on the basis of a wide background of prior experience. Indeed, in this study the amount of "shrinkage" in every case proved to be at most relatively negligible, and in some cases there was an actual increase in the composite-predictor correlation with criterion in the second sample. This contrasts sharply with the substantial shrinkage obtained in the Mosteller and Wallace (1964) study. The reason for this difference is that Mosteller and Wallace were forced to select a small number of predictor variables *during* the study from a much larger set of potential predictors. Thus they capitalized on chance in their selection. Only by cross-validating this selection were they able to obtain an accurate appraisal of the true predictability of their criterion. Their one outstanding predictor variable, however, actually improved on cross validation. This also is not atypical, for if one variable is an outstanding predictor, then it is chosen on its true merit rather than for its error, and hence it can yield either a lower or higher value on cross validation. It is when many variables have uniformly low true correlations that cross validation shrinkage is large. An even more drastic shrinkage occurred in a vintage study reported by Guttman (1941): In this case, the use of 84 regression coefficients in a sample of 136 produced a multiple correlation of 0.73, but when these same weights were used in a second sample of 140, the multiple correlation was 0.04.

Considering each of these groups as a sample from some larger (hypothetical) population, it is clear that the weights obtained in either year can only be approximations to the optimal weights. Hence in using such weights we are not using the true linear regression weights. However, the results of this study (and other studies) suggest that an approximate optimal linear combination of predictor variables often performs nearly as well as the true optimal com-

bination. Geometrically we would say that the composite variable correlation surface is reasonably flat in the region of the point determined by the linear regression weights.

### 13.5 The Effect of Relative Test Lengths on Reliability and Validity: The Multiple Predictor Case\*

As we indicated in Chapter 5, the validity coefficient of any test containing errors of measurement can be increased by increasing the length of the test. Formula (5.11.2) gives the validity of a test at length  $k$  with respect to a fixed criterion in terms of its validity at unit length, its reliability at unit length, and the value  $k$ . Formula (5.10.1) gives the reliability of a test of length  $k$  in terms of its reliability at unit length and the value  $k$ .

Since the multiple correlation coefficient is, in fact, the zero-order correlation between the best linear combination of the predictor variables  $X_1, X_2, \dots, X_n$  and the criterion  $X_0$ , we might suppose the multiple correlation coefficient varies as the lengths of the various predictors are altered. In this section, we shall develop formulas in matrix notation for the effects of changes in test length on reliability, validity, predictor variable intercorrelation, partial regression weights, and multiple correlation. Among other things, we shall show that the partial regression weights depend on the lengths of the various tests, and indeed that the desirability of including a particular variable in a regression equation may depend on the total available testing time.

Let  $X_1, X_2, \dots, X_n$  be a set of  $n$  predictor variables and  $X_0$  be a criterion variable. Let

- $\mathbf{D}_a$  be a diagonal matrix whose diagonal elements are the lengths of the predictors  $X_1, X_2, \dots, X_n$ ,
- $\boldsymbol{\rho}$  be the vector of validity coefficients of  $X_1, X_2, \dots, X_n$  with  $X_0$ ,
- $\mathbf{P}$  (upper case rho) be the matrix of intercorrelations of the predictors,
- $\mathbf{D}_r$  be the diagonal matrix whose diagonal elements are the reliabilities of the predictors  $X_1, X_2, \dots, X_n$ ,
- $\boldsymbol{\beta}$  be the vector of partial regression weights of  $X_1, X_2, \dots, X_n$  with  $X_0$ , and
- $R_a^2$  be the multiple correlation of  $X_1, X_2, \dots, X_n$  with  $X_0$ .

We assume that each of the above quantities is known. Now suppose the length of each of the predictors is altered and the new predictors are denoted by  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ . The length of the criterion is assumed to remain unchanged. Let

- $\mathbf{D}_b$  be a diagonal matrix whose diagonal elements are the new lengths of the predictors  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ .

---

\* Reading of this and the following section may be omitted without loss of continuity.

We assume for present purposes that the new lengths are given. Let

$\tilde{\rho}$  be the vector of validity coefficients of  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$  with  $X_0$ ,

$\tilde{\mathbf{P}}$  be the matrix of intercorrelations of the altered predictors,

$\tilde{\mathbf{D}}_r$  be the diagonal matrix whose diagonal elements are the reliabilities of the altered predictors,

$\tilde{\beta}$  be the vector of partial regression weights of  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$  with  $X_0$ ,

$R_b^2$  be the multiple correlation of  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$  with  $X_0$ .

We assume, for present purposes, that these quantities are unknown. Our task is to express them as functions of the known quantities. For further convenience, let

$\Sigma$  be the dispersion (variance-covariance) matrix of  $X_1, X_2, \dots, X_n$ ,

$\tilde{\Sigma}$  be the dispersion matrix of  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ ,

$\mathbf{D}_\sigma^2$  be Diag  $\Sigma$ , that is, a diagonal matrix whose diagonal elements are the diagonal elements of  $\Sigma$ , namely, the variances of  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ ,

$\tilde{\mathbf{D}}_\sigma^2 = \text{Diag } \tilde{\Sigma}$ ,

$\mathbf{D}_e = \mathbf{D}_b \mathbf{D}_a^{-1}$ ,

$\Lambda = [\mathbf{I} + (\mathbf{D}_e - \mathbf{I}) \mathbf{D}_r] \mathbf{D}_e^{-1}$ .

Then

$$\mathbf{P} = \mathbf{D}_\sigma^{-1} \Sigma \mathbf{D}_\sigma^{-1}$$

and

$$\tilde{\mathbf{P}} = \tilde{\mathbf{D}}_\sigma^{-1} \tilde{\Sigma} \tilde{\mathbf{D}}_\sigma^{-1}.$$

The validity coefficients, intercorrelation matrix, variances, and reliabilities of the predictors at altered length, and the squared multiple correlation and regression weights for the predictors at altered length, are given by

### Theorem 13.5.1

$$\tilde{\rho} = \Lambda^{-1/2} \rho, \quad (13.5.1a)$$

$$\tilde{\mathbf{P}} = \Lambda^{-1/2} (\mathbf{P} + \Lambda - \mathbf{I}) \Lambda^{-1/2}, \quad (13.5.1b)$$

$$\tilde{\mathbf{D}}_\sigma^2 = \Lambda \mathbf{D}_e^2 \mathbf{D}_\sigma^2, \quad (13.5.1c)$$

$$\tilde{\mathbf{D}}_r = \Lambda^{-1} \mathbf{D}_r, \quad (13.5.1d)$$

$$R_b^2 = \rho' (\mathbf{P} + \Lambda - \mathbf{I})^{-1} \rho, \quad (13.5.1e)$$

$$\tilde{\beta} = \mathbf{D}_\sigma^{-1} \mathbf{D}_e^{-1} (\mathbf{P} + \Lambda - \mathbf{I})^{-1} \rho \sigma_0. \quad (13.5.1f)$$

*Proof.* Results (13.5.1a, b, c, d) are straightforward vector and matrix generalizations of the standard formulas for variances, reliabilities, and intercorrelations when the lengths of two tests are subject to variation. Equation

(13.5.1e) is derived as follows:

$$\begin{aligned} R_b^2 &= \tilde{\rho}' \tilde{\mathbf{P}}^{-1} \tilde{\rho} = \rho' \mathbf{\Lambda}^{-1/2} [\mathbf{\Lambda}^{-1/2} (\mathbf{P} + \mathbf{\Lambda} - \mathbf{I}) \mathbf{\Lambda}^{-1/2}]^{-1} \mathbf{\Lambda}^{-1/2} \rho \\ &= \rho' \mathbf{\Lambda}^{-1/2} \mathbf{\Lambda}^{1/2} (\mathbf{P} + \mathbf{\Lambda} - \mathbf{I})^{-1} \mathbf{\Lambda}^{1/2} \mathbf{\Lambda}^{-1/2} \rho = \rho' (\mathbf{P} + \mathbf{\Lambda} - \mathbf{I})^{-1} \rho. \end{aligned}$$

Equation (13.5.1f) is derived from

$$\tilde{\beta} = \tilde{\mathbf{D}}_{\sigma}^{-1} \tilde{\mathbf{P}}^{-1} \tilde{\rho} \sigma_0 = \mathbf{D}_{\sigma}^{-1} \mathbf{D}_e^{-1} (\mathbf{P} + \mathbf{\Lambda} - \mathbf{I})^{-1} \rho \sigma_0. \quad \square$$

The vector  $\tilde{\beta}$  of revised regression weights depends on the matrix  $\mathbf{\Lambda}$  and hence on the diagonal matrix  $\mathbf{D}_b$  of revised test lengths. Also the altered squared multiple correlation may be expressed in terms of the original beta weights and variances instead of the original validities; this result is given in

### Corollary 13.5.2

$$R_b^2 = \beta' \mathbf{D}_{\sigma} \mathbf{P} (\mathbf{P} + \mathbf{\Lambda} - \mathbf{I})^{-1} \mathbf{P} \mathbf{D}_{\sigma} \beta \sigma_0^{-2}. \quad (13.5.2)$$

The proof of this result follows directly from (13.5.1e) and from the standard relation  $\beta = \mathbf{D}_{\sigma}^{-1} \mathbf{P}^{-1} \rho \sigma_0$ .

## 13.6 The Determination of Relative Test Lengths to Maximize the Multiple Correlation

The problem of determining the optimal relative test lengths so as to maximize the predictive validity of a test battery has been studied by Taylor (1939, 1950), by Horst (1949, 1956), and more recently and more generally by Woodbury and Novick (1968). Here we shall follow the development of the last-cited reference.

Suppose  $T$  is the total time available for testing, and a battery  $\mathcal{B}$  of  $n$  predictor variables is available. We assume that we can shorten or lengthen each of the tests of the battery. We wish to determine the amount of time  $t_i$ ,  $i = 1, 2, \dots, n$ , where  $\sum t_i = T$  is fixed, to be allotted to each test so that the multiple correlation between the set of predictors and the criterion is maximized. We omit the proof of the results to be stated here because of its length and because it requires some rather advanced mathematical techniques. It may be found in Woodbury and Novick (1968).

Let  $\rho[X_i(t_i), X'_i(t_i)]$  be the reliability of  $X_i(t_i)$ , the observed score of variable  $i$  at time  $t_i$ , and let

$$a_{ii} = \frac{\sqrt{t_i}}{\sqrt{\sigma^2[X_i(t_i)] \{1 - \rho[X_i(t_i), X'_i(t_i)]\}}}. \quad (13.6.1)$$

Let  $\mathbf{D}$  be a diagonal matrix whose  $ii$ th term is  $a_{ii}$ , and let  $\boldsymbol{\theta}$  be a column vector whose  $i$ th term is  $\sigma[X_i(t_i), Y]$ , where  $Y$  is the criterion variable. Then let

$$\mathbf{F} = \mathbf{D}_t^{-1} (\mathbf{D} \Sigma \mathbf{D} - \mathbf{D}_t) \mathbf{D}_t^{-1} \quad (13.6.2)$$

and

$$\boldsymbol{\gamma} = \mathbf{D}_t^{-1} \mathbf{D} \boldsymbol{\theta} / \sqrt{\sigma_Y^2}. \quad (13.6.3)$$

It may be shown that  $\mathbf{F}$  is the variance-covariance matrix of the true-score variables  $T_i$  when the  $X_i$  have been scaled so that the error-score variables  $E_i$  satisfy  $\mathcal{E}[E_i(1)]^2 = 1$ . It may also be shown that  $\boldsymbol{\gamma}$  is the vector of covariances of the  $T_i$  with the criterion when the  $X_i$  have been scaled so that  $\mathcal{E}[E_i(1)]^2 = 1$  and  $Y$  has been scaled so that  $\mathcal{E}(\eta^2) = 1$ , where  $\eta$  is the true score of the criterion.

Now consider

$$\mathbf{t}^* = \frac{T^* + \mathbf{e}'\mathbf{F}^{-1}\mathbf{e}}{\mathbf{e}'\mathbf{F}^{-1}\boldsymbol{\gamma}} \mathbf{F}^{-1}\boldsymbol{\gamma} - \mathbf{F}^{-1}\mathbf{e}, \quad (13.6.4)$$

where  $T^*$  is the total testing time available and  $\mathbf{e}' = \{1, 1, \dots, 1\}$ , and where each predictor has now been reflected, if necessary, so that its true-score partial regression weight is positive. Woodbury and Novick show that if this equation yields a  $\mathbf{t}^*$  whose elements are all nonnegative, then  $\mathbf{t}^*$  is the vector of optimal test lengths. The squared multiple correlation and vector of regression weights for the given predictor variables at their optimal lengths can then be found from (13.5.1e), (13.5.1f), and the multiple correlation and regression weights obtained from fitting all predictor variables at their given lengths.

If one or more of the elements of  $\mathbf{t}^*$  are negative, then the solution obtained from (13.6.4) is invalid. It is not possible to obtain the correct solution by assigning zero times to these variables and making a proportional adjustment in the other test lengths. However, a simple solution algorithm may be employed to obtain the correct solution.

Given the assumption that the partial correlation of the true score of each predictor variable with criterion is nonzero, it may be shown that if  $T^*$  is sufficiently large, then each predictor should receive a positive time allocation. Thus, if one or more elements of  $\mathbf{t}^*$  are negative, a larger value of  $T^*$  may always be found such that we obtain a solution  $\mathbf{t}^*$  whose elements are all positive. We can do this by reflecting variables as necessary so that each has a positive true-score partial regression weight. Then all of the elements of  $\mathbf{F}^{-1}\boldsymbol{\gamma}$  are positive and we can find the smallest value of  $T^*$  in (13.6.4) for which all elements of  $\mathbf{t}^*$  are nonnegative.

Suppose that we have any valid solution; then we may decrease  $T^*$  until some element of  $\mathbf{t}^*$  is zero. Let us denote this point by  $T_n$ . We can obtain a solution for  $T^* = T_n$ . Then eliminating this variable from the predictor set and proceeding, as in the first instance, we may determine successive values  $T_{n-1}, T_{n-2}, \dots, T_i, \dots, T_2$  at which variables should be excluded and obtain solutions at these points. If we renumber the variables in the reverse of the order in which they drop out of the solution as  $T$  decreases, then the values  $T_1 = 0, T_2, T_3, \dots, T_i, \dots, T_{n-1}, T_n$  are the values of  $T$  at which tests  $i = 1, 2, \dots, n$  should first be used.

We shall illustrate this method by an example given by Taylor, who considers a problem defined by the data in Fig. 13.6.1. Application of the techniques of this section yield an optimal relative assignment of  $t_1^* = 0.308$  and  $t_2^* = 0.692$  for the two variables of the battery when the total available testing time and

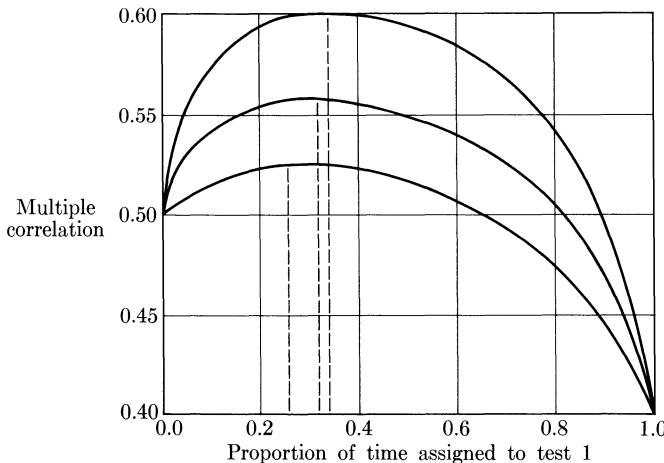


FIG. 13.6.1. Example of the effect of relative test length on multiple correlation. The relative lengths were assigned to two tests of a battery. From top to bottom, the three curves of the figure represent predictor intercorrelations of 0.00, 0.20, and 0.40. The initial lengths of each test are taken as unity, and the total testing time available is also taken as unity. The reliabilities of the tests are  $r_{11} = 0.9$  and  $r_{22} = 0.8$ ; the validities are  $\rho_{01} = 0.4$  and  $\rho_{02} = 0.5$ . [Reprinted from C. W. Taylor, Maximizing predictive efficiency for a fixed total testing time. *Psychometrika*, 1950, 15. Used by permission.]

the initial lengths of each test are taken as unity. Further it is found that for  $T \leq 0.0178$  all available time should be assigned to variable 1. For  $T \geq 0.0178$  the optimal allocation is  $0.0178 + 0.295(T - 0.0178)$  for variable 1 and  $0.0000 + 0.705(T - 0.0178)$  for variable 2. It is of interest to note that for short total testing times variable 1 receives a larger allocation than variable 2; for longer times variable 2 receives the larger allocation.

The middle curve in Fig. 13.6.1 shows the relationship in the Taylor problem between the multiple correlation coefficient and the fraction of the total time devoted to test 1. The maximum value in the middle curve is reached above the base-line value 0.308. Note that the maximum is on the left side of the curve. This fact illustrates the following conclusion: If test 2 is more valid and less reliable than test 1, and if, therefore, proportionately more validity would be lost by shortening test 2 than by shortening test 1, then optimal prediction is to be obtained by shortening test 2. At the extreme left side of the curve the total time is devoted to test 2, and at the extreme right side of the curve the total time is devoted to test 1. At these points, the curves drop to the validities of the individual tests. It may be noted that in the division of time between two tests there can be cases where the combined predictive efficiency is less than it would be if all the time were devoted to the (initially) more valid test.

The two other curves in Fig. 13.6.1 were obtained by taking an intercorrelation of 0.00 in one case and 0.40 in the other case, with all other parameters

**Table 13.6.1**

Intercorrelation, reliabilities (major diagonal elements),  
and validities (ninth column) for time allocation example

	Integration	Induction	Object sorting	Direct memory	Verbal fluency	Reference reading	SAT Verbal	SAT Math	Validity	Time in minutes
Integration	(0.76)	0.25	0.13	0.31	0.28	0.32	0.43	0.43	0.44	20
Induction	0.25	(0.74)	0.07	0.22	0.15	0.25	0.25	0.31	0.17	20
Object sorting	0.13	0.07	(0.82)	0.15	0.09	0.15	0.13	0.16	0.15	20
Direct memory	0.31	0.22	0.15	(0.58)	0.19	0.32	0.46	0.37	0.39	20
Verbal fluency	0.28	0.15	0.09	0.19	(0.70)	0.33	0.33	0.13	0.30	20
Reference reading	0.32	0.25	0.15	0.32	0.33	(0.64)	0.40	0.32	0.40	20
SAT-Verbal	0.43	0.25	0.13	0.46	0.33	0.40	(0.86)	0.44	0.61	75
SAT-Math	0.43	0.31	0.16	0.37	0.13	0.32	0.44	(0.84)	0.39	75

remaining the same as in the sample problem. Of the three cases, the zero intercorrelation case gives the curve with the highest maximum. As the intercorrelation becomes higher, the maximum drops and shifts to the left in an accelerated fashion, having values 0.34, 0.31, and 0.25, respectively, for the intercorrelations of 0.00, 0.20, and 0.40. The reader will certainly wish to note that in this problem each of the curves studied is very flat within a range of at least  $\pm 0.1$  around the optimal relative allocation to variable 1.

We may further illustrate the use of the time allocation algorithm with some data given by John W. French. From the correlations given in Table 13.6.1, the optimal time allocation for a total available time of 150 minutes is (25.6, 0.5, 0.8, 6.8, 0.0, 25.2, 91.1, 0.0) with a multiple correlation of 0.6645. The multiple correlation using SAT-V and SAT-M, each taking 75 minutes, is 0.6248. Dropping all variables except Integration, Reference Reading, and SAT-V yielded a multiple correlation of 0.6555 at the given present lengths. An optimal time allocation of (26.6, 27.3, 96.1) for these variables yielded a multiple correlation of 0.6642. This suggests the use of the allocation (30, 30, 90) with a corresponding multiple correlation of 0.6640, which is just 0.0005 less than the optimal eight-variable solution, yet 0.0392 higher than the multiple correlation using only SAT-V and SAT-M. This increment can be expected to suffer some small shrinkage in cross validation.

### Exercises

- 13.1. Let  $\mathbf{R}$  be the intercorrelation matrix of a set of measurements and  $\mathbf{D}_r$ , a diagonal matrix of the reliabilities of the measurements. Show that the matrix of disattenuated correlations is given by

$$\mathbf{D}_r^{-1/2}(\mathbf{R} - \mathbf{I} + \mathbf{D}_r)\mathbf{D}_r^{-1/2}.$$

*Note:* The matrix  $(\mathbf{R} - \mathbf{I} + \mathbf{D}_r)$  is often called a reduced correlation matrix. It differs from  $\mathbf{R}$  in that its diagonal terms are the correlations between the measurements and parallel forms (the reliabilities) rather than the correlation between the measurement and itself.

13.2. Using the intermediate results from Exercise 13.1 show that the matrix

$$\mathbf{D}_r^{-1/2}(\mathbf{R} - \mathbf{I} - \mathbf{D}_r)$$

is just the matrix  $\mathbf{R} - \mathbf{I} + \mathbf{D}_r$  in which each correlation in the first row has been corrected for attenuation in the first variable, each correlation in the second row has been corrected for attenuation in the second variable, etc. Thus each of the reliabilities is corrected for attenuation in the second subscript. For example,  $\rho_{12}$  is corrected for attenuation with respect to 1, and  $\rho_{21}$  is corrected for attenuation in 2, and so on. But  $\rho_{12} = \rho_{21}$ , and so on, and hence the off-diagonal elements of  $\mathbf{D}_r^{-1/2}(\mathbf{R} - \mathbf{I} - \mathbf{D}_r)$  give all the correlations disattenuated in one variable.

13.3. Show that Eqs. (13.2.6) and (13.2.7) are equivalent.

13.4. Consider a forward variable selection procedure for adding variables to a multiple predictor set and suppose that the predictor variables are numbered to correspond with the order in which they enter this set. Thus, for example,  $\rho_{0 \cdot 123}^2 \geq \rho_{0 \cdot 12k}^2$ ,  $k > 3$ . It would seem reasonable to assume that the increments in the squared multiple correlation would decrease as new variables were added; for example, that

$$\rho_{0 \cdot 123}^2 - \rho_{0 \cdot 12}^2 \geq \rho_{0 \cdot 1234}^2 - \rho_{0 \cdot 123}^2.$$

Using (12.4.8), show that this equation holds if and only if

$$\rho_{04 \cdot 123}^2 \leq \frac{\rho_{03 \cdot 12}^2}{(1 - \rho_{03 \cdot 12}^2)}.$$

- 13.5. Consider the values  $\rho_{04 \cdot 12} = 0.3$ ,  $\rho_{03 \cdot 12} = 0.4$ , and  $\rho_{43 \cdot 12} = -0.7$ . Use Exercise 12.7 to show that these are admissible values. Then show that  $\rho_{04 \cdot 123}^2 = 0.2243$  while  $\rho_{03 \cdot 12}^2 / (1 - \rho_{03 \cdot 12}^2) = 0.1905$  and hence that the first equation is false. Suppose further that  $\rho_{0 \cdot 12}^2 = 0.5$ . Then show that  $\rho_{0 \cdot 123}^2 - \rho_{0 \cdot 12}^2 = 0.08$  while  $\rho_{0 \cdot 1234}^2 - \rho_{0 \cdot 123}^2 = 0.0942$ .
- 13.6. Show that the true partial correlation between  $X_0$  and  $X_1$  with the effect of  $X_2$  partialled out is directly proportional to  $\rho_{01} - \rho_{12}\rho_{02}\rho_{22}^{-1}$ , where  $\rho_{22}$  is the reliability of  $X_2$ . From this show that corresponding observed-score and true-score partial regression weights may have opposite signs.

## References and Selected Readings

- BURKET, G. R., A study of reduced rank models for multiple prediction. *Psychometric Monograph*, 1964, No. 12.
- DRAPER, N. R., and H. SMITH, *Applied regression analysis*. New York: Wiley, 1966.
- Dwyer, P. S., The square-root method and its use in correlation and regression. *Journal of the American Statistical Association*, 1945, **40**, 493–503.
- EFROYMSON, M. A., Multiple regression analysis. In A. Ralston and H. S. Wilf (Eds.), *Mathematical methods for digital computers*. New York: Wiley, 1960.

- EFROYMSON, M. A., Stepwise regression—a backward and forward look. Paper presented at the Eastern Regional Meetings of the Institute of Mathematical Statistics, April 27–29, 1966.
- ELFVING, G., The item-selection problem and experimental design. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961, pp. 81–87. (See also pp. 88–108.)
- ELFVING, G., ROSEDITH SITGREAVES, and H. SOLOMON, Item selection procedures for item variables with a known factor structure. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961. (Also in *Psychometrika*, 1959, **24**, 189–205.)
- FORTIER, J. J., Simultaneous linear prediction. *Psychometrika*, 1966, **31**, 369–381. (a)
- FORTIER, J. J., Simultaneous nonlinear prediction. *Psychometrika*, 1966, **31**, 447–455. (b)
- FRENCH, J. W., The validity of new tests for the performance of college students with high-level aptitude. *Research Bulletin 63–7*. Princeton, N.J.: Educational Testing Service, 1963.
- GUTTMAN, L., Mathematical and tabulation techniques. Supplementary study B. In P. Horst (Ed.), *Prediction of personal adjustment*. New York: Social Science Research Council, 1941.
- HORST, P. (Ed.), *The prediction of personal adjustment*. New York: Social Science Research Council, 1941.
- HORST, P., Determination of optimal test length to maximize the multiple correlation. *Psychometrika*, 1949, **14**, 79–88.
- HORST, P., A note on optimal test length. *Psychometrika*, 1950, **15**, 407–408.
- HORST, P., Optimal test length for maximum differential prediction. *Psychometrika*, 1956, **21**, 51–66.
- HORST, P., *Psychological measurement and prediction*. Belmont, Calif.: Wadsworth, 1966.
- HORST, P., and CHARLOTTE MACEWAN, Predictor elimination techniques for determining multiple prediction batteries. *Psychological Reports*, 1960, **7**, 19–50.
- KENDALL, M. G., *A course in multivariate analysis*. London: Griffin, 1957.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. Vol. II: *Inference and relationship*. London: Griffin, 1961.
- LINHART, H., A criterion for selecting variables in a regression analysis. *Psychometrika*, 1960, **25**, 45–58.
- LORD, F. M., Efficiency of prediction when a regression equation from one sample is used in a new sample. *Research Bulletin 50–40*. Princeton, N.J.: Educational Testing Service, 1950.
- MOSTELLER, F., and D. WALLACE, *Inference and disputed authorship: the Federalist*. Reading, Mass.: Addison-Wesley, 1964.
- NICHOLSON, G. E., Prediction in future samples. In I. Olkin *et al.*, *Contributions to probability and statistics, essays in honor of Harold Hotelling*. Stanford: Stanford University Press, 1960.
- OLKIN, I., and J. W. PRATT, Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 1958, **29**, 201–211.

- ROZEBOOM, W. W., *Foundations of the theory of prediction*. Homewood, Ill.: The Dorsey Press, 1966.
- RYDBERG, S., *Bias in prediction*. Stockholm: Almqvist and Wiksell, 1963.
- STEIN, C., Multiple regression. In I. Olkin *et al.*, *Contributions to probability and statistics, essays in honor of Harold Hotelling*. Stanford: Stanford University Press, 1960.
- SUMMERFIELD, A., and A. LUBIN, A square-root method of selecting a minimum set of variables in multiple regression. *Psychometrika*, 1951, **16**, 271-284.
- SWOYER, V. H., On the best  $k$  of  $n$  predictors. Cambridge: Harvard University, 1966. Unpublished doctoral dissertation.
- TAYLOR, C. W., A method of combining tests into a battery in such a fashion as to maximize the correlation with a given criterion for any fixed total time of testing. Salt Lake City: University of Utah, 1939. Master's thesis.
- TAYLOR, C. W., Maximizing predictive efficiency for a fixed total testing time. *Psychometrika*, 1950, **15**, 391-406.
- TAYLOR, H. C., and J. T. RUSSELL, The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, 1939, **23**, 565-578.
- THORNDIKE, R. L., *Personnel selection*. New York: Wiley, 1949.
- WHERRY, R. J., Appendix A. In W. H. Stead and C. P. Shartle (Eds.), *Occupational counseling techniques*. New York: American Book Company, 1940, pp. 245 ff.
- WOODBURY, M. A., and M. R. Novick, Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. *Journal of Mathematical Psychology*, 1968, **5** (to appear).

## MEASUREMENT PROCEDURES AND ITEM-SCORING FORMULAS

### 14.1 Introduction

The general problem of obtaining the maximum amount of information from a given set of items may be divided into three major components. The first of these is the *measurement procedure*, that is, the manner in which examinees are asked to respond to the individual test items and the conditions under which these responses are elicited. The second component is the specification of the numerical assignment rule, the *item-scoring formula* used for each item. The third component is the combination of item scores into a total test score by means of an *item-weighting formula*. We have already considered the third component briefly in Chapter 12 and we shall consider it in more detail in Chapters 15 and 20. We shall consider the first two components in detail in the present chapter.

In a typical multiple-choice examination, the examinee is presented with a question, offered a number of possible responses, and instructed to choose the one correct response. If he does choose the correct response, either because he knows the correct response or because he guesses successfully, he receives a score of one. If he does not, either because he does not answer the question or because he selects an incorrect response, he receives a score of zero. The examinee's total test score is then taken to be the simple sum of these zero-one item scores.

The major advantages of this measurement procedure, item-scoring formula, and item-weighting formula, are those of efficiency and simplicity for the examinee and the test scorer. More items of this type can be administered in a fixed period of time than items requiring more complicated responses. Also the cost for scoring examination papers from this kind of test is relatively small.

It should be noted that the problems of item scoring and item weighting are usually not independent. Since item weights are typically taken to depend only on the overall characteristics of the items, they could be absorbed into the item score. Basically the item-scoring formula is simply a method for specifying a particular interval-scale quantification for the information from the measurement procedure. For example, we may decide that the responses to a single-choice three-response item should be scored so as to preserve the inherent order of degree of correctness of the three responses. Then the scores  $(1, 0, -1)$  or some strictly increasing function of these values would be an acceptable scoring

formula. We would then determine the exact manner in which these scores should be combined with scores on other items by specifying the item-weighting formula that would specify the weight and possibly a shift for each item score, so that the total test score  $x$  would be given by

$$x = \sum_{i=1}^n (a_i y_i + b_i),$$

where the  $y_i$  are the item scores. Since interval scale properties are preserved under linear transformation, such item weighting does not affect the interval scaling of the individual items. The interval scaling of the total score, however, depends on the weighting constants  $a_i$ , and it changes if these are changed in any way other than multiplying each of them by the same constant.

In evaluating measurement procedures and item-scoring formulas, various kinds of costs must always be kept in mind. For example, it may be that examinees are available for a relatively long period for testing, but that test items are very difficult to obtain. Then we would want to obtain as much information as possible from each item, and hence we would be tempted to employ more complicated measurement procedures, if they were indeed useful.

On the other hand, it may be that items are plentiful but examinee time is scarce. It may also be reasonable to assume that per unit of *time*, we can probably get more information by adding more items (if available) than by introducing complex measurement procedures. Then we would probably be inclined to use the simpler measurement procedure so that we might administer as many items as possible in the limited amount of time.

## 14.2 Guessing and Omitting

Much of mental test theory is designed to deal with dichotomously scored items. The items in almost all cognitive tests, however, permit at least three different responses. For true-false items, for example, the possible responses are "true", "false", and no response, the last being called an *omit*. Multiple-choice items permit several possible responses.

We might suppose that if an examinee knows the answer to an item he encounters, he gives that answer, and that if he does not know the answer, he either omits the item or *guesses* at random. This simple, rather mechanistic model, is rarely accurate, and we consider it here primarily as a prototype of other more realistic models.

Now if we assume the simple *knowledge or random guessing* model, we can easily develop a scoring formula that for some purposes corrects for the independent, additive error introduced by this guessing. However, examinees who have *partial information* about an item do not respond at random, nor do examinees with *misinformation* about that item. In these situations, wrong answers cannot be equally attractive to the examinee. No simple correction formula for guessing is appropriate in these cases.

Neither is there any simple correction for omitting an item. If all examinees were equally likely to guess when their knowledge is incomplete, or equivalently, were equally likely to omit an item, omitting would cause little concern. The fact is that examinees differ widely in their willingness to omit items. This means that the number-right test score (the number of correct answers) usually measures not only the ability of interest but also to some extent the examinee's willingness to omit items—his *omissiveness*. The question of omissiveness is just one aspect of the general problem of response styles to which psychologists and psychometricians are becoming increasingly more sensitive (see Berg, 1967, and papers and references therein).

An obvious approach to the problem of omissiveness is to require each examinee to answer every item in the test. In this case, the number-right test score is no longer contaminated by the examinee's omissiveness, but there is danger that this new requirement has introduced a considerable amount of random error into the test scores. This is certainly true whenever there are examinees who do not have time to study all the test questions, since these examinees are forced to respond at random. The addition of such random error lowers both the reliability and the validity of the test (see Exercise 14.4).

Note that the examinee's true score on multiple-choice and other common types of tests is automatically increased if omitted responses are replaced by random guesses. It is possible to redefine true score (as in Zimmerman and Williams, 1965a, 1965b, 1966) so that it remains unchanged in such a situation, but the true score so defined would not be uncorrelated with the errors of measurement.

Now if we assume the very simple knowledge or random guessing model, then the *chance* score  $x_c$  on any test is defined as the expected score that an examinee would obtain by random guessing on all items. On an  $A$ -choice item the chance score is  $1/A$ ; hence the chance score on a test containing  $n$   $A$ -choice items is

$$x_c = n/A. \quad (14.2.1)$$

Thus, if an examinee always responds at random, his true score (expected score) is necessarily equal to  $x_c$ .

By the usual formula for the variance of a binomial distribution, the sampling variance of the score that an examinee would obtain by random guessing is

$$\text{Var}(X_a) = n(A - 1)/A^2, \quad A = 2, 3, \dots \quad (14.2.2)$$

This variance can be used to define an interval around the chance score within which any score might plausibly have been obtained by random guessing.

It is sometimes argued that differences among scores at and below the chance score should be treated as meaningless. We may easily see that this conclusion is ordinarily incorrect. Suppose that examinee  $a$  knows the answer to  $\kappa_a$  of the  $n$   $A$ -choice items in the test and guesses at random at the others. We may then

determine the likelihood function for his observed score  $x_a$  from the binomial model

$$\begin{aligned} \text{Prob}(X_a = x_a | n, p, \kappa_a) &= \text{Prob}(x_a - \kappa_a | n - \kappa_a, p) \\ &= \binom{n - \kappa_a}{x_a - \kappa_a} p^{x_a - \kappa_a} (1 - p)^{n - x_a}, \\ &\quad x_a \geq \kappa_a, \quad \kappa_a = (0, 1, \dots, n). \end{aligned} \quad (14.2.3)$$

If  $x_a \leq x_c$ , the likelihood function varies with  $x_a$  and with  $\kappa_a$ . This shows that different observed scores below the chance level do provide differential information about  $\kappa_a$ , the level of knowledge of the examinee.

Some formulas and numerical results relating to the effects of random guessing are given by Carroll (1945, 1961), Horst (1932a, 1932b, 1933, 1954a, 1954b), Roberts (1962), and Mattson (1965). Careful empirical studies are reported by Plumlee (1952, 1954); and a small study, by Lord (1964). Some of the references given in the next section are also relevant here.

### 14.3 A Simple Formula Score

A psychometrician might feel that he would like to recover the distribution of scores  $\kappa$  that would have been observed if no random guessing had occurred. If  $g(\kappa)$ ,  $\kappa = 0, 1, \dots, n$ , is the distribution of  $\kappa$  in a population of examinees, and if the assumptions that led to (14.2.3) hold, then the joint distribution of  $X$  and  $\kappa$  is

$$f(x, \kappa) = \binom{n - \kappa}{x - \kappa} p^{x - \kappa} q^{n - x} g(\kappa), \quad x = \kappa, \kappa + 1, \dots, n, \\ \kappa = 0, 1, \dots, n. \quad (14.3.1)$$

The marginal distribution of  $X$  is

$$\varphi(x) = q^{n-x} \sum_{\kappa=0}^n \binom{n - \kappa}{x - \kappa} p^{x - \kappa} g(\kappa), \quad x = 0, 1, \dots, n. \quad (14.3.2)$$

These are a set of  $n + 1$  linear equations that can be solved to express the  $n + 1$  unknown frequencies  $g(0), g(1), \dots, g(n)$  in terms of the values of  $\varphi(x)$ . If  $g(\kappa)$  can be estimated in this or any other way, then the bivariate distribution of  $X$  and  $\kappa$  can be estimated and used in turn to obtain an estimate of the value of  $\kappa$  for each examinee from his observed value of  $X$ .

Unfortunately the psychometrician does not have exact values for  $\varphi(x)$ ,  $x = 0, 1, \dots, n$ , but only the approximations represented by the observed frequencies in the sample of examinees at hand. Substituting these approximations into (14.3.2) is likely to lead to negative estimates for some of the  $g(\kappa)$ —an absurd result. Such negative values can be avoided by linear programming techniques, but the estimated  $g(\kappa)$  is still likely to be intolerably irregular. For the present, no entirely satisfactory methods seem to be available to deal

with this problem. Moreover, since the model itself seems artificial, this line of investigation does not seem promising.

Instead of trying to estimate what the distribution of scores would have been if there had been no random guessing, it is common practice to use a formula score that is an approximation to  $\tilde{\kappa}_a$ , the maximum likelihood estimator of  $\kappa_a$  in (14.2.3). To find the value of  $\kappa_a$  that maximizes (14.2.3), consider for what values of  $\kappa'$

$$\frac{\text{Prob}(X = x \mid \kappa = \kappa')}{\text{Prob}(X = x \mid \kappa = \kappa' + 1)} < 1.$$

Substituting from (14.2.3) and dropping the prime and the subscript  $a$ , we have

$$\frac{\frac{(n - \kappa)!}{(x - \kappa)!(n - x)!} p^{x-\kappa} q^{n-x}}{\frac{(n - \kappa - 1)!}{(x - \kappa - 1)!(n - x)!} p^{x-\kappa-1} q^{n-x}} < 1;$$

or, canceling terms,

$$\frac{n - \kappa}{x - \kappa} p < 1;$$

or, finally,

$$\kappa < (x - np)/q.$$

Thus the likelihood of  $X = x_a$  is maximized when  $\kappa_a = \tilde{\kappa}_a$ , where  $\tilde{\kappa}_a$  is the smallest integer satisfying  $\tilde{\kappa}_a \geq (x_a - np)/q$ . If the examinee omits certain items for which he does not know the answer, the same formula is still valid, provided that  $n$  is replaced by  $n_a$ , the number of items to which examinee  $a$  responds:

$$\tilde{\kappa}_a < (x_a - n_a p)/q. \quad (14.3.3)$$

In common practice, the quantity on the right of (14.3.3), here denoted by  $\hat{\kappa}_a$ , is used as the examinee's *formula score*:

$$\hat{\kappa}_a = \frac{x_a - n_a p}{1 - p} = x_a - \frac{p(n_a - x_a)}{1 - p} = x_a - \frac{w_a}{A - 1}, \quad (14.3.4)$$

where

$$w_a = n_a - x_a \quad (14.3.5)$$

is the number of wrong answers given by the examinee. This formula score ranges from  $\hat{\kappa} = -n_a p/q$  to  $\hat{\kappa} = n_a$ . Different negative values of  $\hat{\kappa}$  have different implications for the examinee and hence supply valid discrimination just as number-right scores below the chance level do, as we remarked near the end of the preceding section. It is established practice to refer to a score obtained from (14.3.4) as an examinee's *formula score*. In this book, we shall use the term *formula score* generically to refer to a score obtained from any specified scoring formula other than the simple zero-one scoring scheme.

#### 14.4 Properties of the Simple Formula Score

By (14.2.3), if the examinee responds to  $n_a$  items, then the expected contribution of guessing to his score is

$$\mathcal{E}'(X_a - \kappa_a) = (n_a - \kappa_a)p,$$

where  $\mathcal{E}'$  denotes a conditional expectation over all randomized guesses with  $\kappa_a$ ,  $p$ , and  $n_a$  fixed. The expected value of the observed score is thus

$$\mathcal{E}'X_a = \kappa_a + (n_a - \kappa_a)/A. \quad (14.4.1)$$

We now find from (14.3.4) that

$$\mathcal{E}'\hat{\kappa}_a = \kappa_a. \quad (14.4.2)$$

*If examinee  $a$  knows the answer to  $\kappa_a$  items and responds at random to  $n_a - \kappa_a$  items, then the formula score  $\hat{\kappa}_a$  is an estimator of  $\kappa_a$  that is unbiased (when  $n_a$  is fixed) over repeated independent random guessing.*

For this reason, the formula score  $\hat{\kappa}_a$  is said to be *corrected for chance success*. The reader should note that the correction is for random guessing only; the correction has no clear validity when the examinee obtains wrong answers by nonrandom guessing or because of misinformation.

*When there are no omitted responses ( $n_a = n$ ), the formula score  $\hat{\kappa}_*$  is perfectly correlated over examinees with the number-right score  $X_*$ .* This important result is evident from the first part of Eq. (14.3.4). Since the origin and scale of measurement is an arbitrary matter, *the correction for chance success has no significant effect on the measurement properties of a test when there are no omitted responses.* However, when there are omitted responses, the formula score for each item is trichotomous  $[1, 0, -1/(A - 1)]$ . Since the number-right score is dichotomous  $(0, 1)$ , the two kinds of scores cannot be perfectly correlated.

One advantage of using the simple formula score and advising all examinees of its use, and of the implications of this use, is a reduction in the variation in omnissiveness among the examinees. One property of the simple formula score is that if it is assumed that the examinee guesses randomly among the possible responses, then his expected formula score on that item is zero. However, if he is able to identify at least one response as being definitely wrong, his expected formula score for guessing among the remaining responses is positive. The reader will easily verify that in a five-choice item, the correct elimination of one, two, three, and four distractors raises the expected guessing score to  $\frac{1}{16}$ ,  $\frac{1}{6}$ ,  $\frac{3}{8}$ , and 1. In general, the expected guessing score after correctly eliminating  $e$  of the  $A$  choices as distractors is

$$(A - e)^{-1} \left( 1 - \frac{A - e - 1}{A - 1} \right). \quad (14.4.3)$$

If the examinee can eliminate at least one distractor as definitely false it will

be to his advantage, on the average, to select one of the remaining possible responses at random. If this is pointed out to the examinee, he will presumably be encouraged to "guess intelligently" on those items on which he has partial information. Experience suggests, however, that students do not respond uniformly to such encouragement.

The error variance of a formula score can be obtained by the general methods (Sections 7.3, 7.5, 8.3, and 8.7) used for the error variance of any test score, or by application of the item-sampling approach of Chapter 11. This error variance can be separated into two additive parts: the part due to guessing, called the *topastic error variance*, and the remainder, called the *scedastic error variance*.

If we consider  $n_a$ , the number of items attempted by examinee  $a$ , to be fixed, we see from (14.2.3) that the topastic error variance of his number-right score is the usual binomial variance,

$$\text{Var}(X_a | n_a, \kappa_a, p) = (n_a - \kappa_a)pq, \quad (14.4.4)$$

the number of trials being  $n_a - \kappa_a$ . By (14.3.4) and (14.4.4), the topastic error variance of the examinee's formula score is

$$\text{Var}(\hat{\kappa}_a | n_a, \kappa_a) = \frac{1}{q^2} \text{Var}' X_a = (n_a - \kappa_a) \frac{p}{q} = \frac{n_a - \kappa_a}{A - 1}. \quad (14.4.5)$$

Many more formulas relating to topastic errors are given by Carroll (1945, 1961) and by Zimmerman and Williams (1965a, 1965b, 1966).

A comparison of (14.4.4) and (14.4.5) shows that the topastic error variance of the formula score  $\hat{\kappa}_a$  is larger than the topastic error variance of the number-right score  $x_a$  by the factor  $1/q^2$ . Can the reader properly interpret this comparison? Actually this factor merely represents the fact that the unit of measurement (and the range) of  $\hat{\kappa}_a$  is  $1/q$  times the unit of measurement and the range of  $X_a$ . This assertion can be made because when  $n_a$  is fixed, these two methods of scoring yield scores that are perfectly correlated.

How does this simple formula scoring affect test reliability? Glass and Wiley (1964) quote empirical results and give a theoretical derivation showing that reliability is lowered by the correction. Their theoretical results are based on the assumption that the number  $gx$ , say, of correct guesses by the examinee is uncorrelated with his true score. This seems to be an implausible assumption. An examinee with a very low true score will tend to have  $gx$  near  $n_a/A$ , whereas an examinee with a true score of  $n$  will necessarily have  $gx = 0$ ; consequently the correlation between  $gx$  and true score will be negative and large in the absence of any counteracting tendency. Lord (1963) has treated a model that avoids this particular objection.

The reader should note that the choice of a scoring method should *not* be determined primarily on the basis of score reliability. Such a procedure is

likely to assign heavy scoring weight to some characteristic that is very reliably measured; whereas the characteristic that can be most reliably measured is not necessarily the characteristic that the examiner wishes to measure. Thurstone (1931) reported that the most reliable weighted composite of  $x_a$  and  $w_a$  for a certain test required that each wrong answer be assigned a positive weight, and not the negative weight to be expected from the reasoning leading to (14.3.4). The reason for this odd result was that there were few omits; thus  $x_a + w_a$  was a good measure of  $n_a$ , the number of items the examinee had time to answer in the limited time allowed. The number  $n_a$  represented the speed of the examinee, and this was the examinee variable that could be measured most reliably by this particular test; the variable representing the examinee's amount of knowledge could not be measured as reliably.

How does this simple formula scoring affect validity? Guilford (1956, 4th edition) has reported gains in validity of 0.02 to 0.03 when the simple formula score (14.3.4) is used, and larger gains when the optimal weighting provided by the multiple regression coefficients for predicting criterion from  $x_a$  and  $w_a$  is used. A theoretical investigation by Lord (1963) indicates that gains in validity can be expected from the use of the simple formula score, but that these gains are usually smaller than those summarized by Guilford. The reader should note that any theoretical treatment such as Lord's is likely to be based on the assumption that examinees are not guided by partial information or by misinformation. This assumption is contrary to fact for most tests, although it is less untrue of some tests than of others. For example, items requiring extensive mathematical derivation may be so written that partial information is of little help to the examinee in finding the correct answer. In Chapter 15, we shall briefly discuss the effects of omitting, guessing, and formula scoring on item statistics.

The simple knowledge or random guessing model is used extensively despite its several weaknesses. One such weakness is that it ignores the possible day-to-day variations in examinee performance which have been the heart of our development of test theory and which experience has taught us cannot be neglected. An equally serious weakness is that the model assumes that if the examinee is unable to pinpoint the correct response, then he is completely ignorant in this situation and has no basis for choosing among the possible responses.

This second assumption can seldom be seriously entertained. Indeed the incorrect responses to an item, the *distractors*, are typically chosen to be widely different in attractiveness. Thus, although the very poorest examinee might need to guess at random on the most difficult item, an examinee who is only slightly better should be able to eliminate at least one distractor on most items; and on the average, a better student should be able to eliminate a greater number of distractors. A somewhat more realistic model might assume random guessing after the elimination of one or more of the distractors, the number eliminated being a (probabilistic) function of the examinee's ability.

### 14.5 A Simple Regression Model for Scoring Items\*

Chernoff (1962) has coupled the assumption of random guessing with the criterion of minimum mean squared error and thereby devised a scoring scheme of some theoretical interest. His work is generically related to an earlier paper by Calandra (1941) and to the standard mean squared error regression formulas of Section 3.7. For this simple model, it is found that an incorrect response is scored zero and a correct response is scored  $\lambda/\pi$ , where  $\pi$  is the proportion of correct responses to the item in the population and

$$\lambda = (A\pi - 1)/(A - 1) \quad (14.5.1)$$

is the "true" proportion of people in the population knowing the correct response to the item,  $A$  being the number of possible responses to the item. The expression for  $\lambda$  is derived by noting that the proportion who answer correctly is given by

$$\pi = \lambda + (1/A)(1 - \lambda).$$

Suppose that each individual who obtains the correct answer receives a score of  $X = x_c$  and the others receive a score of  $X = x_w$ . Now we wish to employ the obtained score as an estimate of the "true" level of knowledge of the examinee on the item, this estimate being one for an individual who knows the answer and zero for an individual who does not.

For the proportion  $\lambda$  who know the answer (and hence respond correctly), the error of estimate is  $x_c - 1$ . For the proportion  $\pi - \lambda$  who do not know the answer but guess correctly, the error of estimate is  $x_c - 0 = x_c$ . And finally, for the proportion  $1 - \pi$  who do not know the answer and guess incorrectly, the error of estimate is  $x_w - 0 = x_w$ . Consequently the mean squared error is given by

$$V = \lambda(x_c - 1)^2 + (\pi - \lambda)x_c^2 + (1 - \pi)x_w^2.$$

Minimizing with respect to  $x_c$  and  $x_w$ , we obtain the scores

$$x_c = \lambda/\pi = (A\pi - 1)/(A - 1)\pi, \quad x_w = 0, \quad (14.5.2)$$

which are the linear regression estimates of the examinee's "true score". The corresponding minimum variance is

$$V_{\min} = \frac{\pi}{\lambda}(\pi - \lambda). \quad (14.5.3)$$

As Chernoff has pointed out, this compares favorably with

$$V^* = \pi - \lambda, \quad (14.5.4)$$

the mean squared error for the usual scoring system, which assigns a one to a correct answer and a zero to an incorrect answer.

---

\* Reading of the remaining sections of this chapter may be omitted without loss of continuity.

If two items are scored jointly, then the responses on each item may be used to estimate each true score in a way that improves the mean squared error of prediction on the other item. This reduction is obtained by using more information, namely, the combined scores. This estimate requires more computation than the scoring system for a single item. It also requires that the model be extended so that it includes an assumption of the independence of guesses for individuals who do not "know" the answers. Had the original random guessing assumption been reasonable, this extension would not be questionable.

Let  $\lambda_{11}$  be the proportion of subjects knowing the answer to both questions,  $\lambda_{10}$  the proportion knowing the answer to the first only,  $\lambda_{01}$  the proportion knowing the answer to the second only, and  $\lambda_{00}$  the proportion knowing the answer to neither. Similarly define  $\pi_{cc}$ ,  $\pi_{cw}$ ,  $\pi_{wc}$ , and  $\pi_{ww}$  in terms of the proportions who answer the questions correctly ( $c$ ) and incorrectly ( $w$ ). Then

$$\begin{aligned}\pi_{cc} &= \lambda_{11} + \frac{\lambda_{10}}{A} + \frac{\lambda_{01}}{A} + \frac{\lambda_{00}}{A^2}, & \pi_{cw} &= \lambda_{10} \frac{(A-1)}{A} + \lambda_{00} \frac{(A-1)}{A^2}, \\ \pi_{wc} &= \lambda_{01} \frac{(A-1)}{A} + \lambda_{00} \frac{(A-1)}{A^2}, & \pi_{ww} &= \lambda_{00} \left( \frac{A-1}{A} \right)^2.\end{aligned}\quad (14.5.5)$$

Let  $x_{1cc}$ ,  $x_{1cw}$ ,  $x_{1wc}$ , and  $x_{1ww}$  be the scores assigned for the  $i$ th question to individuals whose results for the pair of questions are  $cc$ ,  $cw$ ,  $wc$ , and  $ww$ , respectively,  $i = 1, 2$ . Then one may obtain the minimum mean squared error scores by minimizing

$$\begin{aligned}V_1 &= \left( \lambda_{11} + \frac{\lambda_{10}}{A} \right) (x_{1cc} - 1)^2 + \left( \frac{\lambda_{01}}{A} + \frac{\lambda_{00}}{A^2} \right) x_{1cc}^2 + \frac{(A-1)\lambda_{10}}{A} (x_{1cw} - 1)^2 \\ &\quad + \lambda_{00} \frac{(A-1)}{A^2} x_{1cw}^2 + \left[ \lambda_{01} \frac{(A-1)}{A} + \lambda_{00} \frac{(A-1)}{A^2} \right] x_{1wc}^2 \\ &\quad + \lambda_{00} \left( \frac{A-1}{A} \right)^2 x_{1ww}^2; \\ V_2 &= \left( \lambda_{11} + \frac{\lambda_{01}}{A} \right) (x_{2cc} - 1)^2 + \left( \frac{\lambda_{10}}{A} + \frac{\lambda_{00}}{A^2} \right) x_{2cc}^2 + \frac{(A-1)}{A} \lambda_{01} (x_{2wc} - 1)^2 \\ &\quad + \lambda_{00} \frac{(A-1)}{A^2} x_{2wc}^2 + \left[ \lambda_{10} \frac{(A-1)}{A} + \lambda_{00} \frac{(A-1)}{A^2} \right] x_{2cw}^2 \\ &\quad + \lambda_{00} \left( \frac{A-1}{A} \right)^2 x_{2ww}^2.\end{aligned}\quad (14.5.6)$$

The optimal scores are found to be

$$\begin{aligned}x_{1wc} &= x_{1ww} = x_{2cw} = x_{2ww} = 0, \\ x_{1cc} &= \frac{\lambda_{11} + (\lambda_{10}/A)}{\pi_{cc}}, & x_{2cc} &= \frac{\lambda_{11} + (\lambda_{01}/A)}{\pi_{cc}}, \\ x_{1cw} &= \frac{(A-1)\lambda_{10}/A}{\pi_{cw}}, & x_{2wc} &= \frac{(A-1)\lambda_{01}/A}{\pi_{wc}}.\end{aligned}\quad (14.5.7)$$

The reader may obtain the minimum values of  $V_1$  and  $V_2$  and show that each is smaller than the corresponding residual variance from the one-variable model (see Exercise 14.1).

There is an interesting and perhaps paradoxical feature of this model. Suppose that two questions are of unequal difficulty ( $\lambda_1 > \lambda_2$ , say) and that the total score is just the sum of the item scores. Then the subject who answers the easier question correctly and the more difficult question incorrectly will have a higher total score than the subject who misses the easier question but answers the more difficult question correctly.

In theory, this two-variable approach can be extended to any number of variables. Chernoff, however, points out that this extension is not very practical because of computational difficulties and the enormous increase in the number of subjects needed to estimate the appropriate  $\lambda$ 's reliably. It should also be pointed out that this theory provides no rationale for combining the item scores into a test score.

#### **14.6 The Regression Method with a Simple Model That Assumes Partial Knowledge**

The unrealistic *knowledge or random guessing* model of the previous section can be replaced by a second model originally suggested by Horst and called by Chernoff the *simple ordering of responses* model. This model assumes first that for each item it is possible to order (and number) the  $A$  responses uniquely, such that *every* examinee's knowledge will consist of knowing that the correct response is among the first  $r$  responses from some  $r$ . Thus an examinee's true score on an item will lie between 1 and  $A$ . This model further assumes that the examinee, in responding, selects one of these  $r$  responses at random. Thus the examinee, on the basis of knowledge, rules out certain distractors and then guesses randomly. Although this model involves the unrealistic assumption that the order in which distractors are ruled out is the same for all examinees, it does begin to consider questions that a realistic model cannot ignore.

Let  $\gamma_j$  be the proportion of the population who can limit their choice to the first  $j$  responses, and  $\pi_j$  the proportion who select the  $j$ th response. Then

$$\pi_j = \frac{\gamma_j}{j} + \frac{\gamma_{j+1}}{j+1} + \cdots + \frac{\gamma_r}{r}, \quad 1 \leq j \leq r. \quad (14.6.1)$$

It is easy to verify by substitution that

$$\gamma_j = j(\pi_j - \pi_{j+1}), \quad 1 \leq j \leq r, \quad (14.6.2)$$

where  $\pi_{r+1}$  is taken to be zero. In particular,  $\gamma_1 = \pi_1 - \pi_2$ , that is, the percentage who know the answer is the difference between the observed proportions for the two most favored responses. Let  $c_r$ ,  $r = 1, 2, \dots, A$ , be the true level of an examinee's knowledge corresponding to his ability to eliminate all but the

first  $r$  responses. If partial knowledge is not recognized, the numerical assignment

$$c_1 = 1 \quad \text{and} \quad c_l = 0, \quad l = 2, 3, \dots, A,$$

would be appropriate. If partial knowledge is recognized, some other assignment  $1 = c_1 \geq c_2 \geq \dots \geq c_r = 0$  is appropriate.

Chernoff shows that for this model the minimum mean squared error score for an examinee selecting response  $j$  is

$$x_j = \pi_j^{-1} \sum_{l=j}^A \frac{\gamma_i c_i}{i}. \quad (14.6.3)$$

For the special case  $c_1 = 1, c_l = 0, l = 2, 3, \dots, n$ , this reduces to

$$x_1 = \gamma_1 / \pi_1 = (\pi_1 - \pi_2) / \pi_1. \quad (14.6.4)$$

This model is more realistic than the one of the previous section because it recognizes the possibility that a subject has partial knowledge; the assumption that the responses can be ordered for each item as required above, however, is hardly plausible. Nevertheless the reader interested in working in this area will want to study Chernoff's mathematical techniques very closely.

The following remarks, taken from the last paragraph of Chernoff's paper, are of interest in connection with the appraisal of the relevance of decision theory to mental test theory given in the Introduction:

The validity of a test plays an important role in most studies of item analysis. In this paper we have ignored this question and there is no essential need to introduce it. The fundamental issue here concerns how to estimate what the test does measure. The question of whether the test measures what we would like it to can be treated separately . . .

#### 14.7 Other Item-Scoring Formulas

The possible item-scoring formulas that might reasonably be suggested for use in scoring multiple-choice items seem to be unlimited in number. We shall not attempt to survey many of them here; rather we shall briefly mention just one such formula and the experimental evidence available for evaluating its usefulness.

If an item contains a right or best answer, it would seem compelling to assign the highest score (possibly one) to this response and some lower scores (possibly zero) to the other possible responses. In the unlikely event that an item analysis were to uncover an item that leads, when so scored, to a negative association with the ability being measured or with a criterion of interest, the test constructor would probably be inclined to omit this item from the test.

If we wish to recognize the possibility of partial information or perhaps of misinformation, then we can assign different scores to the various incorrect responses. For example, one distractor might be designed to ferret out common

misinformation. We might call such a distractor, which is literally the least correct response, the *worst distractor*. A possible scoring scheme might assign a score of one to the correct response and a score of  $-s$  to a worst distractor response, where  $0 < s \leq 1$ . Lord (unpublished study) investigated just such a scoring scheme and found that he gained little by using such a scoring formula. Gulliksen (1950) has discussed a number of other reasonable item-scoring formulas, but in no case did he find that using such methods appreciably increases validity. Papers of Cronbach and Merwin (1955), Merwin (1959), and Brown (1965) are also relevant.

#### 14.8 An Evaluation of Partial Knowledge

Despite the long history of disappointing results that has characterized the study of formula scoring, Shuford, Albert, and Massengill (1966) argue that

Upon reflection it is quite apparent that all techniques in current use for assessing the present state of a student's knowledge fail to extract all of the potentially available information. In the case of objective testing . . . the response methods upon which they are based extract only a very small fraction of the information potentially available from each query.

Although the amount of residual information that can in fact be recovered by introducing more refined response methods is subject to question and, much more importantly, to experimental study, the appeal of the suggested possibility of such recovery is unlikely to be ignored. For this reason, we shall devote the remainder of this chapter to considering some new scoring schemes and the mathematical and statistical models that have been proposed to justify them.

Before proceeding, we hasten to advise the reader again that what little experimental work has been done on the traditional methods of formula scoring has not been encouraging, and that no experimental work has been published that supports the new methods. Thus, at present, the sole recommendation of these new methods is their strong conceptual attractiveness. In evaluating any new response method, it will be necessary to show that it adds more relevant ability variation to the system than error variation, and that any such relative increase in information retrieved is worth the effort, in the sense outlined in Section 14.1.

The enticing feature of these new methods is that they involve actual changes in the measurement procedures, that is, the manner in which subjects are required to respond. One such procedure, studied by Coombs, Milholland, and Womer (1956), requires that the examinee mark out all those responses that he believes to be "wrong". Formally, though possibly not psychologically, this procedure is equivalent to the one discussed earlier by Dressel and Schmid (1953) in which the examinee is required to select those choices which *could* be correct.

The item-scoring formula used by Coombs *et al.* assumes a four-choice item with one correct response and three distractors, and gives a +1 for each dis-

tractor correctly marked and a  $-3$  for each correct response marked as a distractor; thus item scores range from  $+3$  to  $-3$ . If an examinee decides at random whether or not to mark a choice as a distractor, then his expected score on that choice is zero. Also each of the scores from  $+3$  to  $-3$  represents a different response, indicating a different amount of knowledge.

An "advantage" of this scoring scheme is that it often tends to discourage guessing. Suppose, for example, an examinee has marked two distractors he knows to be wrong and hence, he believes, has a sure two points to his credit. There remain two choices, one the answer and the other a distractor. If he has no knowledge and he guesses, he is gambling an additional point credit against a three-point loss on what he perceives to be a 50-50 chance. *If the examinee's goal is to maximize his expected score*, then "this is not a profitable game to play". On the other hand, if he feels that he wants to "risk all" in hopes of "beating out" some better students, guessing may indeed be a good strategy. An empirical test of this item-scoring formula yielded a somewhat higher reliability than for the conventional zero-one scoring, but almost no gain in validity.

#### **14.9 Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item**

De Finetti (1965) has recently suggested an appealing general approach to the assessment of partial knowledge. This approach derives from his own work (1964), and that of Savage (1954) and others in the personal probability approach to Bayesian statistical inference. However, it is important to point out that in the context of the assessment of partial knowledge, the question of the validity of a theory of personal probability in no way suggests, or is suggested by, its validity or lack of validity as a model for statistical inference. It is perfectly possible to accept or reject personal probability either as a basis for scientific inference or as a theory that is potentially useful for the evaluation of partial knowledge, and still reject or accept it for the other purpose. As with all other mental test theories, validity of this theory must be established by using it to make and verify important predictions. If the theory of personal probability in application to the assessment of partial knowledge suggests certain measurement procedures and related item-scoring and item-weighting formulas that are then empirically established to be valid predictors, then this theory will have been validated *for this particular purpose*.

The theory of personal probability assumes that for each  $A$ -choice item, the degree of partial knowledge of an examinee relevant to that item can be expressed completely and uniquely by a set of values  $p_j$ ,  $j = 1, 2, \dots, A$ , such that

$$p_j \geq 0, \quad \text{and} \quad \sum_{j=1}^A p_j = 1.$$

The values  $p_j$  are the examinee's personal probabilities that the  $j$ th choice is

the correct one. For  $A = 2$ , the probability space is the line connecting the points  $(1, 0)$  and  $(0, 1)$ . For  $A = 3$ , the probability space consists of the equilateral triangle and its interior formed by connecting the points  $(1, 0, 0)$ ,  $(0, 1, 0)$ , and  $(0, 0, 1)$ , since the triangle and its interior is the locus of all points satisfying the conditions on the  $p_j$  given above. For an  $A$ -choice item, the probability space consists of an  $A$ -simplex. An  $A$ -simplex is that portion of the  $(A - 1)$ -hyperplane,  $p_1 + p_2 + \dots + p_A = 1$  in  $A$ -space, whose coordinates are all nonnegative.

De Finetti considers a number of possible response methods designed to locate the examinee's personal probabilities in some subset of the  $A$ -simplex. The most powerful of these methods, the continuous method, allows the exact identification of each of the probabilities.

A scoring method, for de Finetti, is a numerical assignment that "obliges the [examinee] to reveal his true beliefs, because any falsification will turn out to be to his disadvantage". Let us illustrate this method by example before stating the requirements in general terms.

Suppose the examinee is required to respond with a set of values (weights)  $r_1, r_2, \dots, r_A$ , with  $r_j \geq 0$  and  $\sum r_j = 1$ . Consider the item-scoring formula  $s'$ , where

$$0 \leq s' \equiv r_h + \frac{1}{2}(1 - \sum r_j^2) \leq 1 \quad (14.9.1)$$

and where  $h$  is the correct choice. The maximum score is attained when the total weight is concentrated on the right choice, that is, when the examinee responds  $r_h = 1$ ,  $r_j = 0$  for  $j \neq h$ . The minimum score is attained when the total weight is concentrated on a single wrong choice.

Now consider the expected value of the score  $s'$  with respect to the examinee's personal probability. It may be shown that this expected value is maximized when  $r_j = p_j$  for  $j = 1, 2, \dots, A$ . Suppose then that this scoring formula is used and suppose that all examinees understand its properties, know their  $p$  values, and are able to make the necessary computations. Then, if each examinee is motivated to respond so as to maximize his expected score with respect to his own personal probability distribution, this continuous scoring scheme will actually cause each examinee to reveal that distribution for each item. We shall give some general results on continuous scoring methods in Section 14.11.

Recognizing possible difficulties in requiring examinees to make, without error, the very precise kinds of probability evaluations required by the continuous scoring method, de Finetti has surveyed a number of less demanding procedures and provided scoring schemes for each. He first assumed that the examinee desires to maximize his expected score with respect to his personal probability distribution, and then designed each scoring scheme in such a way as to motivate the examinee to respond consistently with his personal probabilities. In each of these cases, the method determines a proper subset of the  $A$ -simplex in which the examinee's personal probabilities lie, and also provides a score for the examinee on each given item.

For example, consider the very simple instruction, *mark one alternative as correct*. An appropriate scoring scheme assigns values  $s_1$  and  $s_2$ ,  $s_1 > s_2$ , for a correct and an incorrect response, respectively; for example, one for a correct response and zero for an incorrect response. In general, if the examinee selects choice  $j$ , it is then known that for him  $p_j \geq A^{-1}$  and  $p_j \geq p_l$  for all  $l \neq j$ . Thus, from de Finetti's point of view, this very simple measurement procedure yields very little information concerning an examinee's knowledge of this item. Shuford's argument, given at the beginning of Section 14.8, would seem to be strengthened by this analysis.

De Finetti discusses four general classes of measurement procedure that permit partial identification of an examinee's personal probabilities. The first of these contains the *purely rank-order methods*. The procedure based on the instruction to mark one alternative is a trivial member of this class. An entire subclass may be characterized by the instruction, *mark as correct exactly  $A'$  alternatives, cross out exactly  $A''$* , with  $A' + A'' \leq A$ . The *complete ordering method* is the most general purely rank-order method. This method asks the examinee to designate the alternatives believed to be *most, second, ..., least* probably correct by  $1, 2, \dots, A$ , respectively. The different possible scores are only  $A$  in number,  $s_l$  being the score if the correct answer is given rank  $l$  ( $l = 1, 2, \dots, A$ ), with  $s_l$  decreasing as  $l$  increases ( $s_1 > s_2 > \dots > s_A$ ). The complete ordering method gives little information about the examinee's personal probabilities. It may be concluded only that  $A^{-1} \leq p_j < 1$  for  $j = 1$  and  $0 \leq p_j \leq j^{-1}$  for  $j > 1$ , where  $p_j$  is the personal probability of the  $j$ th-ranked choice. Generally the ordering methods give rather poor bounds on the personal probabilities.

The second class is *flexible schemes with two or three permitted levels of response*. Coombs' method described earlier, when generalized to an  $A$ -choice item, is a typical member of this class. This class generally comprises methods characterized by the instruction, *mark (with a double or single emphasis) one or none, cross out freely*. For the Coombs method, de Finetti suggests the more general scoring formula

$$s_k^+ = k \quad \text{and} \quad s_k^- = k(1 - p), \quad 0 < p \leq 1$$

to obtain the scores for crossing out  $k$  alternatives that do and do not, respectively, include the right answer. In this case, it is to the examinee's advantage to cross out any alternative if the associated personal probability is less than the selected  $p$ . The value  $p$  may be chosen by the examiner.

The third class is that of *strict least distance methods*. Least distance methods are those derived as simplified versions of the continuous method. The most interesting of these is the method *five stars or none*. In this method, a star is used to indicate a unit of personal probability equal to 0.20, and the examinee is instructed to distribute the stars among the alternatives so as to reflect his personal probabilities. If the examinee is indifferent about the alternatives, he

is permitted not to make an assignment; but if he does make an assignment, then he must use all five stars. The examinee's response may then be scored according to Table 14.9.1, where the scores have been reduced to the scale (0, 25) so that the examiner need use only small integers and no fractions. After glancing at this table, the reader will appreciate the complexity for the examinee of the scoring scheme for the five stars or none measurement procedure.

**Table 14.9.1**Scoring system for the *five stars or none* measurement procedure

Type	Number of stars on correct response	Score	Type	Number of stars on correct response	Score
5	5	25	3/1/1	1	12
	0	0		0	7
4/1	4	24	2/2/1	2	18
	1	8		1	13
	0	4		0	8
3/2	3	21	2/1/1/1	2	19
	2	16		1	19
	0	6		0	9
3/1/1	3	22	1/1/1/1/1	1	15
	0	0		0	10

De Finetti recognizes that this scoring formula is complicated and cannot be taught to the examinee. However, he remarks that the "intuitive" meaning of the possibility of distributing the probability as  $0.6 + 0.4$  or  $0.6 + 0.2 + 0.2$ , for example, and an understanding of the idea of a better agreement of one's preferences with one distribution rather than another are the only requirements presupposed, and that these can probably be improved by training. De Finetti also suggests that this kind of training may be generally beneficial to the examinee.

The fourth and final class of noncontinuous procedures are those referred to as *least distance methods (flexible)*. These methods are basically like those of the third class except that certain restrictions are lifted. For example, the method *five stars used flexibly* is the same as *five stars or none* except that the examinee has the option of using as few or as many of the five stars as he chooses. Little can be said here about the value of these procedures. We might remark, though, that any small gain in flexibility is counterbalanced by a very large increase in complexity of the scoring formula.

#### 14.10 Assumptions Underlying the Personal Probability Approach to Item Scoring

If we take validity here to mean the recovery of an examinee's partial knowledge and not necessarily predictive validity in the sense of Chapter 12, we can observe that the validity of de Finetti's scoring schemes depends on three very demanding assumptions.

1. The scoring method, as well as the permitted modes of responding, must be known to the subjects. Furthermore subjects must not only know the method but learn to understand fully its implications with particular reference to behavior in the face of uncertainty. Finally they must be able to make the necessary computations to determine an optimal strategy for each item.
2. Examinees must be keenly interested in obtaining a high total score, precisely in the sense of maximizing their total expected score.
3. They must be able to assign numerical values to their subjective probabilities accurately and reliably.

For most of de Finetti's proposed procedures, the first assumption cannot possibly be satisfied, and for many it does not seem that it can be even reasonably approximated. Although examinees may feel comfortable with and may even "like" a response method such as *five stars or none*, it is hardly likely that many examinees will be able to keep track of all of the implications of the various possible responses (see Table 14.9.1). At best it might be possible to explain to the examinees the intricacies of the continuous response method and then introduce *five stars or none* as an approximation. The continuous scoring procedure itself possibly is not subject to this criticism since the examinees need only be convinced that it is in their best interest to respond honestly.

The second assumption could prove even less satisfactory than the first. In some testing situations, examinees ought not try to maximize their total *expected* score. Such a strategy might well be a very bad strategy for some students. The problem here is the one which is *consistently* encountered whenever an attempt is made to apply decision theoretic methods to the analysis of human behavior. For such applications, there is seldom reason to suppose that we know each subject's utilities. In the present situation,

*there is little reason to assume that utility is linear with expected score for every examinee.*

The fault lies not in utility theory, which is perfectly general and perfectly correct, but in an inappropriate application of this theory.

Suppose a very large group of examinees take a scholarship examination and it is well known that only those scoring in the top 10% on the test will be awarded scholarships. A student who believes that his knowledge places him

in the upper 10% might then rationally choose a strategy which tends to maximize his total expected score because such a strategy would probably come close to maximizing the probability of his score being in the top 10%. Actually it is only this second maximization that has relevance to the problem. A less able student, however, may well be able to find a strategy that yields a somewhat lower expected total score but, because of a higher variance in the distribution of his possible total scores, yields a higher personal probability that his total score will place him in the upper 10%. In general, the less able an examinee believes himself to be, the riskier a strategy he *ought* to take in a situation such as this. In many situations, this means that a good student should not do very much random guessing but that a poor student should, even if there is a penalty for guessing. De Finetti's suggestion that such behavior can be repressed by making the test sufficiently long is generally correct in theory, but it needs to be established that it will be possible in practice to make tests sufficiently long for this purpose. Also any such lengthening is self-defeating since the whole rationale for considering these measurement procedures is that they extract more information per item and hence require fewer items to attain a specified precision.

Over and above questions of how examinees *ought* to behave in such situations, there are even more pertinent questions of how they *do* behave. What little experimental evidence there is suggests that there are indeed great individual differences in the kind of risk-taking strategy adopted by subjects. A brief survey of "Risk taking and intellective functioning on objective tests" is included in the definitive essay on risk taking by Kogan and Wallach (1967). Adams and Adams (1961) have provided some evidence that subjects respond to training in realism of confidence judgements.

Despite the problems arising from the assumptions required by the personal probability approach to item scoring, this approach does presently represent a very promising line of investigation into methods of extracting more information per test item. One reason for this is that we can now look forward to the direct use of computers in the testing process. With examinees sitting at individual computer consoles on which they directly record their responses, the present clerical expenses of processing complicated responses may be significantly lower. However, with computerized testing a second and perhaps more attractive new approach to testing becomes feasible, namely, the sequential selection of items, such that each examinee is administered a test composed of items that at each stage of the testing are chosen so as to contribute as much as possible to the estimation or decision problem at hand.

The assumptions of the personal probability model are certainly more realistic than the assumptions of the random guessing model. Many of the questions raised in previous paragraphs may well be answered satisfactorily by empirical studies. And even if some of these assumptions are violated, it is very possible that the proposed measurement procedures may be robust with respect to these assumptions. However, as we indicated in the introduction to this chapter,

the evaluation of the value of additional information, if it can be obtained, must be weighed against the added costs in time, effort, and money necessary to obtain it. Furthermore such evaluation must be in concrete terms, such as those of predictive validity, though not necessarily in terms of a validity coefficient.

Finally we should note that although de Finetti provides a rationale for scoring individual items, he provides no rationale for combining these item scores into total test scores. There is no basis for thinking that a simple sum of such item scores would in any sense provide an optimal test score. Indeed the item score in this system is actually only a by-product of a procedure designed primarily to estimate personal probabilities.

### 14.11 Reproducing Scoring Systems

Shuford, Albert, and Massengill (1966) have catalogued some useful results on the existence of appropriate scoring schemes for the continuous response method. Let  $p_j$ ,  $j = 1, 2, \dots, A$ , with  $0 \leq p_j \leq 1$  and  $\sum_{j=1}^A p_j = 1$ , be an examinee's personal probabilities for the  $A$  alternatives of a multiple-choice item. Let the examinee's responses to the alternatives be  $r_j$ ,  $j = 1, 2, \dots, A$ , with  $0 \leq r_j \leq 1$  and  $\sum_{j=1}^A r_j = 1$ . Let  $\mathbf{r} = \{r_1, r_2, \dots, r_A\}$  and  $\mathbf{p} = \{p_1, p_2, \dots, p_A\}$ .

The examinee receives a score  $\varphi_h(\mathbf{r})$ , where  $h$  is the correct answer. Then the examinee's expected score with respect to his own personal probability distribution is

$$\lambda(\mathbf{r}, \mathbf{p}; \varphi_h) = \sum_{h=1}^A p_h \varphi_h(\mathbf{r}). \quad (14.11.1)$$

The question then is: Under what conditions on the scoring system  $\varphi_h(\mathbf{r})$  will an examinee who desires to maximize his expected score with respect to his personal probability distribution be impelled to respond with the allocation  $r_1 = p_1, r_2 = p_2, \dots, r_A = p_A$ , that is, with the allocation  $\mathbf{r} \equiv \mathbf{p}$ ? A scoring system that satisfies these conditions is called a *reproducing scoring system* (RSS). Formally we have

**Definition 14.11.1.** The scoring system  $\varphi_h(\mathbf{r})$  is an RSS if and only if, with respect to variation in  $\mathbf{r}$ ,

$$\max \lambda(\mathbf{r}, \mathbf{p}; \varphi_h) = \lambda(\mathbf{p}, \mathbf{p}; \varphi_h). \quad (14.11.2)$$

Consider the special case of a two-alternative item ( $A = 2$ ). The examinee's personal probabilities may be represented by  $\mathbf{p} = (p, 1 - p)$ , and his response by  $\mathbf{r} = (r, 1 - r)$ . It seems reasonable to consider only scoring schemes that reward accurate response. Thus the higher the probability assigned by the examinee to the correct response, the larger should be the student's score. Let us denote the scoring formulas by

$$\varphi_1(\mathbf{r}) \equiv \varphi_1(r, 1 - r) \quad \text{and} \quad \varphi_2(\mathbf{r}) = \varphi_2(1 - r, r). \quad (14.11.3)$$

It seems sufficient to consider scoring systems having the property that  $\varphi_h(r)$  is a differentiable, nondecreasing function of  $r$ . Denoting the derivatives with respect to  $r$  of  $\varphi_1(r)$  and  $\varphi_2(r)$  by  $\varphi'_1(r)$  and  $\varphi'_2(r)$ , respectively, we have

**Theorem 14.11.2.** If  $A = 2$  and  $\varphi'_1(r)$  and  $\varphi'_2(r)$  are differentiable, non-decreasing functions of  $r$ , then a necessary and sufficient condition that  $\varphi_h(r)$  be an RSS is that

$$p\varphi'_1(p) = (1-p)\varphi'_2(1-p) \quad \text{for all } 0 \leq p \leq 1. \quad (14.11.4)$$

We omit the proof, which is quite straightforward, although the proof of sufficiency must consider the cases  $p = 0$ ,  $p = 1$ , and  $0 < p < 1$  separately.

No restriction has yet been placed on the range of the score  $\varphi$ . It is clear, however, that a scoring formula that assigns a score  $\varphi_1(0) = -\infty$  would be unacceptable since, when combined with other item scores to construct a total test score, this score would overwhelm all others. Thus it is *essential* that the scoring formula be bounded.

It is sometimes possible to construct an unlimited number of bounded reproducing scoring systems. The method of construction is given by

**Theorem 14.11.3.** Let  $h(t)$  be any nonnegative function with a uniformly bounded derivative for  $0 \leq t \leq 1$ , and with  $h(1) = 0$ . Let  $A = 2$ , and let

$$\varphi_1(r, 1-r) = \int_0^r h(t) dt \quad \text{and} \quad \varphi_2(1-r, r) = \int_{1-r}^1 \frac{t}{1-t} h(t) dt. \quad (14.11.5)$$

Then  $\varphi_h(r)$  is a *bounded* reproducing scoring system.

An undesirable feature of the construction of an RSS by the method of Theorem 14.11.3 is that this construction permits an undesirable asymmetry to exist between the scores received by the examinee in that the item score will depend on whether the correct answer is numbered one or two. Under most circumstances, it seems reasonable to demand that the score received by an examinee "who assigns probability  $r$  to the correctness of an answer arbitrarily designated 1 is the same as the score he would receive if he assigned a probability  $r$  to the correctness of the same answer arbitrarily designated 2". This requirement leads to

**Definition 14.11.4.** If  $A = 2$  and  $\varphi_h(r)$  is a reproducing scoring system, and if

$$\varphi_1(r, 1-r) = \varphi_2(1-r, r), \quad (14.11.6)$$

then  $\varphi_h(r)$  is a symmetric reproducing scoring system (SRSS).

Two symmetric reproducing scoring systems for binary items are of particular interest. The first of these is given by the formula

$$\varphi(r) = 1 - (1-r)^2, \quad 0 \leq r \leq 1. \quad (14.11.7)$$

This is a version of the “quadratic” scoring system studied previously by de Finetti (1964), van Naerssen (1961), and others. The second is given by the formula

$$\varphi(r) = \frac{r}{[r^2 + (1 - r)^2]^{1/2}}, \quad 0 \leq r \leq 1, \quad (14.11.8)$$

and referred to as the two-dimensional “spherical” scoring system.

A definition of an SRSS for items with more than two alternatives can be given as a direct generalization from the two-alternative case. This definition is given by Shuford *et al.* (1966), who have also given  $A$ -alternative SRSS’s that are analogous to the two-alternative “quadratic” and “spherical” scoring systems. We shall not be able to consider these details here.

Formula scores that depend on the entire set  $\mathbf{r}$  of an examinee’s responses are undesirable for two reasons. First, it is difficult for the examinee to appreciate clearly the implications of the scoring formula. Second, the cost of obtaining such a formula score for each item could well be prohibitive. What would be desirable and perhaps essential is a symmetric reproducing scoring system based on a score that depends only on the response to the correct alternative, and not on the responses to the incorrect alternatives.

Unfortunately, when the number of possible responses exceeds two, the only scoring system satisfying this requirement is the logarithmic formula score (Shuford *et al.*, 1966)

$$\varphi(r) = a \log br, \quad a > 0, \quad b > 0. \quad (14.11.9)$$

This formula score is unacceptable since we can hardly assign an examinee a score of minus infinity when he makes the response  $r = 0$  to the correct alternative. It is puzzling that such an apparently rational approach to the problem should yield a unique but obviously unacceptable solution.

Shuford *et al.* (1966) suggest the following truncated logarithmic scoring function:

$$-1 \leq \varphi(r) \equiv \begin{cases} 1 + \log_{10} r & \text{for } 0.01 < r \leq 1 \\ -1 & \text{for } 0 < r \leq 0.01 \end{cases} \leq 1, \quad (14.11.10)$$

where  $\varphi$  takes the extreme values  $-1$  and  $+1$  when  $r = 0$  and  $r = 1$ , respectively. They do not report any details of their study, but claim that

In brief, this truncation has no effect on the choice of  $r$  when  $.027 \leq p \leq .973$ . This truncated logarithmic scoring system, although not strictly an RSS, does have the reproducing property for values of  $p$  between  $.027$  and  $.973$ . When  $p \leq .027$  ( $p \geq .973$ ), the student’s expected score is maximized by setting  $r = 0$  ( $r = 1$ ).

If some experimental controls on individual response styles can be devised, an empirical testing of this measurement procedure and item-scoring formula would seem to be warranted, even though this method does not strictly satisfy the theory from which it was developed. At least as promising, however, is de Finetti’s five stars or none method.

### Exercises

- 14.1. Show that the expected value of the scoring formula (14.9.1) with respect to the examinee's personal probability distribution is

$$\sum_{j=1}^A r_j p_j + \frac{1}{2} \left( 1 - \sum_{j=1}^A r_j^2 \right).$$

- 14.2. Show that this is maximized when  $r_j = p_j$ .

- 14.3. Give a heuristic resolution to the "paradoxical" property of the Chernoff model.

- 14.4. Show that if  $\rho(X_1, X_0) > 0$  and  $\rho(X_2, X_0) > 0$ , we have  $\rho(X_1 + X_2, X_0) < \rho(X_1, X_0)$  if (and only if)

$$\rho(X_2, X_0) < \rho(X_1, X_0) \frac{\sigma(X_1 + X_2) - \sigma(X_1)}{\sigma(X_2)}.$$

Further, show that  $[\sigma(X_1 + X_2) - \sigma(X_1)]/\sigma(X_2) \leq 1$  with equality if and only if  $\rho(X_1, X_2) = 1$ . Finally, show that it is possible to satisfy the first inequality and still have  $\|\sigma_{ij}\|$ ,  $i = 0, 1, 2, \dots$ , as positive and definite, provided that  $\sigma^2(X_2)$  is sufficiently large.

### References and Selected Readings

- ADAMS, J. K., and PAULINE A. ADAMS, Realism of confidence judgements. *Psychological Review*, 1961, **68**, 33–45.
- BERG, J. A. (Ed.), *Response set in personality assessment*. Chicago: Aldine Publishing Co., 1967.
- BROWN, J., Multiple response evaluation of discrimination. *British Journal of Mathematical and Statistical Psychology*, 1965, **18**, 125–137.
- CALANDRA, A., Scoring formulas and probability considerations. *Psychometrika*, 1941, **6**, 1–9.
- CARROLL, J. B., The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 1945, **10**, 1–19.
- CARROLL, J. B., The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 1961, **26**, 347–372.
- CHERNOFF, H., The scoring of multiple choice questionnaires. *Annals of Mathematical Statistics*, 1962, **33**, 375–393.
- COOMBS, C. H., On the use of objective examinations. *Educational and Psychological Measurement*, 1953, **13**, 308–310.
- COOMBS, C. H., J. E. MILHOLLAND, and J. F. B. WOMER, The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, **16**, 13–37.
- CRONBACH, L. J., and J. C. MERWIN, A model for studying the validity of multiple-choice tests. *Educational and Psychological Measurement*, 1955, **15**, 337–352.
- DAVIS, F. B., Use of correction for chance success in test scoring. *Journal of Educational Research*, 1958, **52**, 279–280.

- DAVIS, F. B., Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, 1959, **19**, 291-298.
- DAVIS, F. B., and G. FIFER, The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, **19**, 159-170.
- DE FINETTI, B., Foresight: its logical laws, its subjective sources. In H. E. Kyburg and H. E. Smokler (Eds.), *Studies in subjective probability*. New York: Wiley, 1964, pp. 93-158.
- DE FINETTI, B., Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 1965, **18**, 87-123.
- DRESSEL, P. L., and P. SCHMID, Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, 1953, **13**, 574-595.
- GLASS, G. V., and D. E. WILEY, Formula scoring and test reliability. *Journal of Educational Measurement*, 1964, **1**, 43-47.
- GUILFORD, J. P., *Fundamental statistics in psychology and education*, 4th ed. New York: McGraw-Hill, 1956.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.
- HORST, P., The chance element in the multiple choice test item. *Journal of General Psychology*, 1932, **6**, 209-211. (a)
- HORST, P., The difficulty of multiple choice test item alternatives. *Journal of Experimental Psychology*, 1932, **15**, 469-472. (b)
- HORST, P., The difficulty of a multiple-choice test item. *Journal of Educational Psychology*, 1933, **24**, 229-232.
- HORST, P., The estimation of immediate retest reliability. *Educational and Psychological Measurement*, 1954, **14**, 705-708. (a)
- HORST, P., The maximum expected correlation between two multiple-choice tests. *Psychometrika*, 1954, **19**, 291-296. (b)
- KOGAN, N., and M. A. WALLACH, Risk taking as a function of the situation. *New Directions in Psychology III*. New York: Holt, Rinehart & Winston, 1967.
- LORD, F. M., Formula scoring and validity. *Educational and Psychological Measurement*, 1963, **23**, 663-672.
- LORD, F. M., The effect of random guessing on test validity. *Educational and Psychological Measurement*, 1964, **24**, 745-747.
- MATTSON, D., The effects of guessing on the standard error of measurement and the reliability of test scores. *Educational and Psychological Measurement*, 1965, **25**, 727-730.
- MERWIN, J. C., Rational and mathematical relationship of six scoring procedures applicable to three-choice items. *Journal of Educational Psychology*, 1959, **50**, 153-161.
- MICHAEL, W. B., R. STEWART, B. DOUGLASS, and J. H. RAINWATER, An experimental determination of the optimal scoring formula for a highly-speeded test under different instructions regarding scoring penalties. *Educational and Psychological Measurement*, 1963, **23**, 83-99.

- NEDELSKY, L., Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement*, 1954, **14**, 459-472.
- PLUMLEE, LYNNETTE B., The effect of difficulty and chance success on item-test correlations and test reliability. *Psychometrika*, 1952, **17**, 69-86.
- PLUMLEE, LYNNETTE B., The predicted and observed effect of chance success on multiple-choice test validity. *Psychometrika*, 1954, **19**, 65-70.
- ROBERTS, A. O. H., The maximum reliability of a multiple-choice test. *Psychologia Africana*, 1962, **9**, 286-293.
- SAVAGE, L. J., *The foundations of statistics*. New York: Wiley, 1954.
- SHERIFFS, A. C., and D. S. BOOMER, Who is penalized by the penalty for guessing? *Journal of Educational Psychology*, 1954, **45**, 81-90.
- SHUFORD, E. H., A. ALBERT, and H. E. MASSENGILL, Admissible probability measurement procedures. *Psychometrika*, 1966, **31**, 125-145.
- SLAKTER, M. J., Risk taking on objective examinations. *Journal of the American Educational Research Association*, 1967, **4**, 31-43.
- STALNAKER, J. M., Weighting questions in the essay-type examination. *Journal of Educational Psychology*, 1938, **29**, 481-490.
- SWINEFORD, F., and P. M. MILLER, Effects of directions regarding guessing on item statistics on a multiple choice vocabulary test. *Journal of Educational Psychology*, 1953, **44**, 129-139.
- THURSTONE, L. L., *The reliability and validity of tests*. Ann Arbor, Mich.: Edwards Brothers, 1931.
- VAN NAERSSEN, R. F., A scale for the measurement of subjective probability. *Acta Psychologica*, 1961, 159-166.
- VOTAW, D. F., The effect of do-not-guess directions on the validity of true-false or multiple-choice tests. *Journal of Educational Psychology*, 1936, **27**, 698-703.
- WILKS, S. S., Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 1938, **3**, 23-40.
- ZILLER, R. C., A measure of the gambling response-set in objective tests. *Psychometrika*, 1957, **22**, 289-292.
- ZIMMERMAN, D. W., and R. H. WILLIAMS, Effect of chance success due to guessing on error of measurement in multiple-choice tests. *Psychological Reports*, 1965, **16**, 1193-1196. (a)
- ZIMMERMAN, D. W., and R. H. WILLIAMS, Chance success due to guessing and non-independence of true scores and error scores in multiple-choice tests: computer trials with prepared distributions. *Psychological Reports*, 1965, **17**, 159-165. (b)
- ZIMMERMAN, D. W., and R. H. WILLIAMS, Interpretation of the standard error of measurement when true scores and error scores on mental tests are not independent. *Psychological Reports*, 1966, **19**, 611-617.

# ITEM PARAMETERS AND TEST CONSTRUCTION

## 15.1 Introduction

The *test item* is the unit building block from which the composite test is constructed. As we pointed out in our discussion of (4.6.6) and (4.6.7), the score  $X$  on most composite tests can be thought of as the sum of the scores on the individual test items  $Y_g$ . Throughout the present chapter, we shall assume this simple case; that is, we shall assume

$$x_a = \sum_{g=1}^n y_{ga}. \quad (15.1.1)$$

In classical item analysis, the first concern is to obtain a description of the statistical characteristics or parameters of the individual test item. However, the individual item is of interest only through its effect on the total test score. Also the statistical characteristics of the total test depend entirely on the statistical characteristics of the items used to build it. Thus the real importance of item analysis arises from the effects of the individual item characteristics on the characteristics of the entire measuring instrument. Knowledge of the item characteristics and their effects helps us understand and allow for the peculiar measurement properties of a particular psychological test. Item analysis may enable us to construct tests with specified or, in a limited sense, optimal measurement properties. Also item parameters are necessary to build a new test form that is as nearly parallel as possible to a given test form.

If a test is composed of items so organically related to each other that the characteristics of each item change substantially whenever it is administered apart from the other items, then, for purposes relating to this test, there is little reason to analyze the items individually. This fact points out a basic assumption of item analysis—the assumption that certain statistical characteristics of the item remain unchanged (to a practical approximation) throughout the changing contexts in which the psychometrician uses it. Without some kind of constancy under change, there would be little point in determining the characteristics of the item at all.

If a test does contain organically related items, it is often possible to arrange these in mutually unrelated groups and to treat each group as a single item for purposes of item analysis. For example, consider the following two questions:

1. What metal is obtained from bauxite?
2. What properties make this metal very useful?

The second question cannot be administered apart from the first. If the two questions cannot be reworded, then it may be desirable to think of them as constituting a single item on which the examinee's score may be either 0, 1, or 2.

A multitude of different item parameters have been proposed for studying and characterizing individual test items. By what criteria does one select a desirable item parameter?

In mental test theory, *the basic requirement of an item parameter is that it have a definite (preferably a clear and simple) relationship to some interesting total-test-score parameter.* This criterion by itself rules out of consideration many of the parameters proposed in the literature. It is desirable, in addition, that an item parameter be estimable by an item statistic that has the properties usually desirable in any statistic, such as small sampling error, easily determined sampling distribution, and convenience of computation.

Item statistics are frequently computed from data collected in a pretest, preliminarily to the assembly of the final test form. It must be noted that the item statistic or parameter describes not only the test item but usually the group of examinees to which the item is administered as well. In practical work, groups of examinees often differ systematically from pretesting to regular testing, or from one regular testing to the next. For this reason, *it is important, if possible, to consider item parameters that remain invariant from one group of examinees to another, at least over certain types of systematic change from group to group.*

The following sections are concerned with item statistics or parameters that have a definite relation to interesting total test-score statistics or parameters and that have, when possible, convenient invariance properties over different groups of examinees.

## 15.2 Item Difficulty

If we take the expectation of both sides of (15.1.1), it is apparent that the test mean score for a group of examinees is equal to the sum over items of the *item mean score*:

$$\mu_X = \sum_{g=1}^n \pi_g, \quad (15.2.1)$$

where

$$\pi_g = \mathcal{E}_a Y_{ga}. \quad (15.2.2)$$

This important relation has already been given as (4.3.4). A parallel statement is valid for sample means:

$$\bar{x} = \sum_{g=1}^n p_g, \quad (15.2.3)$$

where  $\bar{x}$  is the average test score and  $p_g \equiv \sum_a y_{ga}/N$  is the average score on item  $g$  in the sample. For theoretical work, it is usually desirable to consider  $p_g$  an estimate, unbiased over sampling of examinees, for the population item difficulty  $\pi_g$ .

Clearly the mean item score  $\pi_g$  is an important item parameter, satisfying the criterion that it have a definite and simple relationship to an interesting total test-score parameter. *If a test constructor is to select items from a pool of pretested items to produce an n-item test such that a group of examinees similar to the pretest group has a predetermined average score  $\bar{x}$ , then he usually should select items that have an average pretest sample difficulty of  $\bar{x}/n$ .*

### 15.3 Item Discriminating Power

Consider next the variance of the test scores for a population of examinees (all variances, covariances, and correlations in this chapter are taken over examinees, not over items or replications). By the formula for the variance of a sum (4.3.7),

$$\sigma_X^2 = \sigma^2(\sum_g Y_{g*}) = \sum_g \sum_h \sigma(Y_{g*}, Y_{h*}) = \sum_g \sum_h \sigma_g \sigma_h \rho_{gh}, \quad (15.3.1)$$

where  $\sigma_g^2 \equiv \mathcal{E}(Y_{g*} - \mathcal{E}Y_{g*})^2$  is the variance of the scores on item  $g$ , and  $\rho_{gh}$  is the product moment correlation between item  $g$  and item  $h$ . Similarly, for a sample of examinees,

$$s_x^2 = \sum_g \sum_h s_g s_h r_{gh}, \quad (15.3.2)$$

where  $s_x^2 \equiv \sum_a (x_a - \bar{x})^2/N$  and  $s_g^2 = \sum_a (y_{ga} - p_g)^2/N$ . Clearly  $\sigma_g^2$ , the variance for item  $g$ , is an important item parameter having a definite and simple relationship to the variance of the test scores.

In the case of binary (zero-one) items, the variance of the item can be written as a function of the item difficulty, as shown in (4.6.9):

$$\sigma_g^2 = \pi_g - \pi_g^2. \quad (15.3.3)$$

Similarly, for a sample,

$$s_g^2 = p_g - p_g^2. \quad (15.3.4)$$

For binary items, the item difficulty  $\pi_g$  provides all the information provided by  $\sigma_g^2$  and more (because when item  $h$  has  $\pi_h = 1 - \pi_g$ , then  $\sigma_h^2 = \sigma_g^2$ ).

Test items with high  $s_g^2$  are often chosen in preference to others when a test is built from pretested items. The reason is that an item having relatively low  $s_g^2$  compared with other items cannot possibly add much to the variance of an unweighted sum of item scores. For binary items,  $s_g^2$  is largest when  $p_g = \frac{1}{2}$ . Items with  $p_g$  near zero or one have low variance and thus can contribute little to overall test-score variance; such items are little used except for special purposes.

The interitem correlation  $\rho_{gh}$  could be called an important item parameter except for the fact that it cannot be used to predict the properties of a test that contains item  $g$  but not item  $h$ . Item intercorrelations will frequently be of concern, as evidenced by (15.3.1), but they are not basic parameters for describing a single test item. We discuss related parameters for a single item below and in Section 15.7.

Choosing items with high intercorrelations increases the total-score variance. Such items are desirable so long as they measure the psychological trait or traits in question at the time. Sometimes high item intercorrelations arise from a trait of no current interest to the test writer. We discuss the complications that this fact introduces into item selection in the next section.

By treating  $\sigma_X^2$  as the covariance between  $X_* \equiv \sum_g Y_{g*}$  and itself and by using a standard formula (Eq. 4.3.18) for a covariance involving a sum, we obtain a different formula for the variance of the test scores:

$$\sigma_X^2 = \sigma(X_*, \sum_g Y_{g*}) = \sum_g \sigma(X_*, Y_{g*}) = \sum_g \sigma_X \sigma_g \rho_{gX},$$

where  $\rho_{gX}$  is the *item-test correlation*, that is, the product moment correlation between the score on item  $g$  and the score on the total test. If we divide both sides by  $\sigma_X$ , we obtain

$$\sigma_X = \sum_{g=1}^n \sigma_g \rho_{gX}. \quad (15.3.5)$$

Similarly

$$s_x = \sum_{g=1}^n s_g r_{gx}. \quad (15.3.6)$$

*Equations (15.3.5) and (15.3.6) are basic formulas expressing the population and sample standard deviations of test scores in terms of item parameters and item statistics, respectively.*

From the general formula (4.7.3) for a correlation involving a sum, it may be found that

$$\rho_{gX} \equiv \rho(Y_{g*}, \sum_h Y_{h*}) = \frac{1}{\sigma_X} \sum_{h=1}^n \sigma_h \rho_{gh}. \quad (15.3.7)$$

Thus the item-test correlation is a weighted sum of the interitem correlations  $\rho_{gh}$ .

The item-test correlation for item  $g$  varies, of course, if the group of examinees or the nature of test  $X$  is changed. In many applications, however, approximately parallel forms of a test are built year after year and administered to quite similar groups of examinees. In such situations, the  $\rho_{gX}$  for a particular item may be almost the same regardless of the test form in which the item appears. For this reason,  $\rho_{gX}$  is widely used as an item parameter.

Suppose one wishes to build a form  $B$  of a test as nearly parallel as possible to an existing form  $A$ . It is common practice to try to include items in form  $B$  that are matched as closely as possible with the items in form  $A$  on item type, subject matter, difficulty, and item-test correlation. If this is done carefully, one usually obtains a very closely parallel test form.

Given approximations to the item difficulty (or item mean score), the item variance, and the item-test correlation for each item proposed for inclusion in a projected test, (15.2.3) and (15.3.6) enable the test writer, under favorable

conditions, to predict the mean and variance of the test before assembling it. In addition, the specific reliability of the test can be bounded from below by (4.4.4), which we restate here in terms of item parameters:

$$\alpha \equiv \frac{n}{n-1} \left[ 1 - \frac{\sum_{g=1}^n \sigma_g^2}{\left( \sum_{g=1}^n \sigma_g \rho_{gX} \right)^2} \right]. \quad (15.3.8)$$

It is obvious from (15.3.8) that for fixed  $\sigma_g^2$ , the higher the item-test correlations, the higher the coefficient  $\alpha$ . It is common practice to try to choose items with high  $\rho_{gX}$ . Since item  $g$  helps to determine the score  $X$ , the correlation  $\rho_{gX}$  is spurious in a certain sense. An appropriate correction is discussed by Wolf (1967) and by Cureton (1966).

*Item discriminating power* is an imprecise term for the effectiveness of an item in discriminating among “good” and “poor” examinees. The parameter  $\rho_{gX}$  is a rough index of item discriminating power. We shall discuss other indexes of item discriminating power in Sections 16.5 and 16.10, in Chapters 17 through 20, and especially in Section 20.4.

The correlation between two parallel forms can be expressed in terms of item parameters. By using the general formula (4.7.3) for the correlation between two sums, we find that

$$\begin{aligned} \rho(X, X') &= \rho\left(\sum_g Y_{g*}, \sum_{g'} Y_{g'*}\right) \\ &= \frac{1}{\sigma_X \sigma_{X'}} \sum_{g=1}^n \sum_{g'=1}^n \sigma_g \sigma_{g'} \rho_{gg'}. \end{aligned} \quad (15.3.9)$$

As a slight digression, consider the following question: If the total score  $X$  is a weighted sum of binary item scores

$$X_a \equiv \sum_{g=1}^n w_g U_{ga},$$

what choice of weights  $w_g$  maximizes coefficient  $\alpha$ ? [Note that for the weighted case, the formula needed for  $\alpha$  is (4.4.8), not (15.3.8).] The required weights are given in Lord (1958). Here we note briefly that the required weights are the factor loadings of the standardized item scores on their first principal component. These are essentially the same as Guttman's principal components for the weighting system, which, however, are more general in that they are available for polytomous items (Guttman, 1941, pp. 327ff.; Stouffer, 1950, 315–321).

A problem of some importance is that of selecting items to measure changes in examinees occurring between two successive testings (see Saupe, 1966). This is an area in which further work is needed.

### 15.4 Item Validity

Consider next the test validity coefficient  $\rho_{X\nu}$ , the population product moment correlation between the test score  $X$  and an external criterion  $\nu$ . A consistent estimate of this validity coefficient is the corresponding sample correlation  $r_{x\nu}$ .

By the formula for a covariance involving a sum, the covariance between test score and criterion may be written

$$\sigma_{X\nu} = \sigma(\nu, \sum_g Y_g) = \sum_g \sigma_g \sigma_\nu \rho_{g\nu}, \quad (15.4.1)$$

where  $\rho_{g\nu}$  is the item-criterion correlation, or *item validity*, in the sample of examinees; the other symbols have obvious meanings. By (15.3.5) and (15.4.1), the validity coefficient for the test is

$$\rho_{X\nu} = \frac{\sigma_{X\nu}}{\sigma_X \sigma_\nu} = \frac{\sum_g \sigma_g \rho_{g\nu}}{\sum_g \sigma_g \rho_{gX}}. \quad (15.4.2)$$

If the score units are chosen so that  $X$  and  $\nu$  have equal variances, this result may be written

$$\rho_{X\nu} = \bar{\sigma}_{g\nu} / \bar{\sigma}_{gX}, \quad (15.4.3)$$

where  $\bar{\sigma}$  denotes a covariance averaged over all  $g = 1, 2, \dots, n$  items.

*Equation (15.4.2) is a basic formula expressing the test validity coefficient in terms of item parameters.* Thus it is clear that  $\rho_{g\nu}$ , the item validity, is an important item parameter.

A common problem of test design is to select from a pool of pretested items the one particular subset that constitutes the most valid test for predicting a given criterion. The test writer who wishes to do this is faced with a paradox. On the one hand, it has been shown (Eq. 3.9.8) that the square root of the test reliability is an upper bound on the test validity; thus it is common practice to choose items with  $r_{gx}$  as large as possible so as to increase the score variance (15.3.6) and the test reliability (15.3.8). On the other hand, it appears from (15.4.3) that if other things remain unchanged, validity is maximized by making the item-test correlations as small as possible.

It is a fact that among items of equal validity and equal variance, those with the lowest intercorrelations are best. If they have negative intercorrelations, so much the better. This statement can be verified from the following formula for the validity of a composite test, obtained by substituting (15.3.7) into the denominator of (15.4.2), using (15.3.1), and simplifying:

$$\rho_{X\nu} = \frac{\sum_g \sigma_g \rho_{g\nu}}{\sqrt{\sum_g \sum_h \sigma_g \sigma_h \rho_{gh}}}. \quad (15.4.4)$$

We can easily see from this formula that if the validity is positive, then, other things being equal, the lower (or more negative) the item intercorrelations, the higher the validity.

To maximize validity, one wants highly valid items. In practice, however, it is mathematically impossible to find large numbers of items that correlate highly with a criterion and not at all (or negatively) with each other. This may be verified from (15.4.4): Since the test validity is a correlation coefficient and therefore can be no greater than one, it follows from (15.4.4) that

$$\sqrt{\sum_g \sum_h \sigma_g \sigma_h \rho_{gh}} \geq \sum_g \sigma_g \rho_{gv}. \quad (15.4.5)$$

When the right-hand side is positive, as we would ordinarily expect it to be, we may square both sides of (15.4.5) and show that

$$\sum_g \sum_h \sigma_g \sigma_h \rho_{gh} \geq \left( \sum_g \sigma_g \rho_{gv} \right)^2. \quad (15.4.6)$$

Some of the  $\rho_{gh}$  on the left could be negative or zero in (15.4.6) but it is not possible for all of them to be so. A transparently clear illustration is provided by the special case where all items are statistically equivalent and where (15.4.6) therefore becomes

$$\rho_{gh} \geq \rho_{gv}^2.$$

Gulliksen (1950, Section 21.8) has given a convenient graphic method for improving the validity of a test by eliminating some of the test items, a method that has been elaborated by Green (1954). This method is based on the sample analog of (15.4.3),

$$r_{xv} = \bar{s}_{gv}/\bar{s}_{gx}, \quad (15.4.7)$$

where

$$\bar{s}_{gv} = \frac{1}{n} \sum_{g=1}^n s_{gv}, \quad (15.4.8)$$

$$\bar{s}_{gx} = \frac{1}{n} \sum_{g=1}^n s_{gx}.$$

Figure 15.4.1 shows a plot of  $s_{gv}$  against  $s_{gx}$  for  $n = 48$  items from an experimental Law School Admissions Test, the criterion  $v$  being law-school grades.\* The sample validity coefficient is  $r_{xv} = 0.436$ . The centroid of this plot, the point  $(\bar{s}_{gx}, \bar{s}_{gv})$ , is indicated by a cross. (The *centroid* of any distribution is the point whose coordinates are the means of the variables.) We see from (15.4.7) that the sample validity coefficient of the test is the slope of the vector from the origin through the centroid.

---

\* These Educational Testing Service data were kindly made available by Marjorie A. Olsen and W. B. Schrader.

Graphically it is evident that the vector from the origin through the centroid, and thus also the sample validity coefficient, will be raised by eliminating items from the bottom and lower-right edges of the plot. This is precisely the practical procedure recommended. The dashed lines drawn on the plot are contour lines for incremental validity. For a given contour line, the validity of the test will be increased by a constant amount when any single item falling along that line is added to the test.

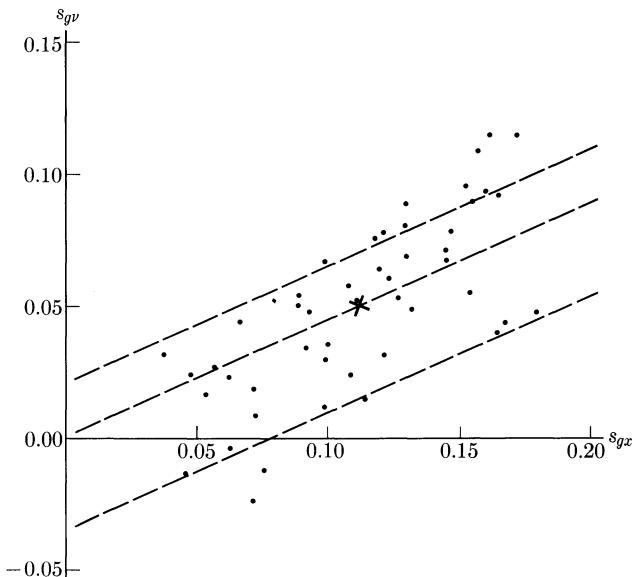


FIG. 15.4.1. Item-criterion covariance plotted against item-test covariance for 48 Law School Admissions Test experimental items.

However, it is important to note that (15.4.7) holds only when the  $x$  in  $r_{gv}$  is the same as the  $x$  in  $s_{gx}$ . To know the sample validity coefficient of the score  $x'$  on a shortened test, we should have  $s_{gx'}$  instead of  $s_{gx}$ . In general,  $s_{gx'} \neq s_{gx}$  and therefore the method recommended cannot yield optimal results. On the other hand, if only a few items are eliminated, then it is likely that  $s_{gx'}$  will not differ too much from  $s_{gx}$ . In this case the method should yield good results, provided that sampling errors in the  $s_{gx}$ -values are sufficiently small.

Any method that selects items on the basis of sample statistics is likely to "capitalize on chance". Items selected because they have high  $r_{gv}$  tend to have  $\rho_{gv} < r_{gv}$ . The effect of this on the validity of a test that has been built by such selection cannot in practice be determined from data on the same sample of examinees used to obtain the  $r_{gv}$ . The effect can be determined from data on a new sample, however, by *cross validation* (see Section 13.2). Satisfactory methods for predicting the degree of shrinkage in cross validation are not

currently available. Shrinkage is greatest when

- 1) the sampling errors of  $r_{gv}$  are large,
- 2) the proportion of items selected is small, and
- 3) the population values  $\rho_{gv}$  are homogeneous.

There is no easy and yet rigorous way of choosing a subset of  $n$  items from a large pool of items so as to maximize test validity (15.4.4). Webster (1956) has contributed a practicable method with some intuitive appeal. The interested reader may refer to the contributions of Darlington and Bishop (1966), Elfving, Sitgreaves, and Solomon (1959, 1961), Elfving (1961a, b, c), Linhart (1960), Horst and MacEwan (1960), Lev (1956), Summerfield and Lubin (1951), and Horst (1934, 1949), among others.

## 15.5 Product Moment Correlations for Dichotomous Items

The correlation between a dichotomous variable and a nondichotomous variable is called a *point biserial correlation*. If the second variable is also dichotomous, the correlation is called a *phi coefficient*. The formulas for the ordinary product moment correlation simplify in these two special cases. The simplified formulas are presented in this section.

Given that  $Y_g$ , the score on item  $g$ , has just two possible values,  $Y_g = \alpha$  and  $Y_g = \beta$ , let us define the binary variable

$$U_g \equiv \frac{Y_g - \alpha}{\beta - \alpha} = 0 \quad \text{or} \quad 1. \quad (15.5.1)$$

Since  $U_g$  is a linear transformation of  $Y_g$ , the correlation of any variable with  $Y_g$  is the same as the correlation with  $U_g$ . Because of this fact, correlation formulas obtained for binary items apply to all dichotomous items, however they may be scored.

The covariance between  $U_g$  and any variable  $v$  is

$$\begin{aligned} \sigma_{gv} &\equiv \mathcal{E}(U_g - \pi_g)(v - \mu) = \mathcal{E}U_g v - \mu \mathcal{E}U_g - \pi_g \mathcal{E}v + \pi_g \mu \\ &= \pi_g \mu^+ - \mu \pi_g, \end{aligned} \quad (15.5.2)$$

where  $\mathcal{E}U_g = \pi_g$  and  $\mathcal{E}v = \mu$ , and where

$$\mu^+ \equiv \mathcal{E}U_g v / \pi_g = \mathcal{E}(v | U_g = 1) \quad (15.5.3)$$

is the mean value of  $v$  for those examinees who answer the item correctly, that is, who have  $U_g = 1$ . Thus, by (15.5.2) and (15.3.3), the product moment correlation between  $U_g$  and  $v$  can be written

$$\rho_{gv} = \frac{(\mu^+ - \mu)\pi_g}{\sigma_v \sigma_g} = \frac{\mu^+ - \mu}{\sigma_v} \sqrt{\frac{\pi_g}{1 - \pi_g}}. \quad (15.5.4)$$

This formula may be rewritten in various ways, for example,

$$\rho_{gv} = \frac{\mu^+ - \mu^-}{\sigma_v} \sqrt{\pi_g(1 - \pi_g)}, \quad (15.5.5)$$

where  $\mu^- \equiv \mathcal{E}(v | U_g = 0)$ . This product moment correlation between a dichotomous variable and another variable is the point biserial correlation coefficient.

We obtain the corresponding *sample point biserial correlation* by replacing  $\mu^+$ ,  $\mu^-$ ,  $\sigma_v$ , and  $\pi_g$  in (15.5.5) by the corresponding sample statistics  $M^+$ ,  $M^-$ ,  $s_v$ , and  $p_g$ ; for example,

$$r_{gv} = \frac{M^+ - M^-}{s_v} \sqrt{p_g q_g}. \quad (15.5.6)$$

Here

$$M^+ \equiv \frac{1}{N p_g} \sum_{a=1}^N u_{ga} v_a = \frac{1}{N p_g} \sum_a^+ v_a, \quad (15.5.7)$$

where  $\sum_a^+$  denotes summation over only those examinees who answer the item correctly.

Das Gupta (1960) has given a careful discussion of the statistical properties of the sample point biserial correlation, which includes a large-sample standard error for  $r_{gv}$  and a variance-stabilizing transformation. Some other writers have treated the point biserial under the assumption that the conditional distribution of  $v$  is normal for  $u = 0$  and for  $u = 1$ , but this is not appropriate for item analysis work, where the unconditional distribution of  $v$  remains the same from item to item while the conditional distributions vary according to the dichotomization.

By similar manipulations, we find that for any two dichotomous variables  $Y_i$  and  $Y_j$ , the usual formula for a product moment correlation can be written simply as

$$\rho_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\sigma_i \sigma_j}, \quad (15.5.8)$$

where  $\pi_{ij}$  is the proportion of cases in the “successful” category for both variables at once. Such a correlation is called a *phi coefficient*. The *sample phi coefficient* is, similarly,

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i q_i p_j q_j}}, \quad (15.5.9)$$

where  $\rho_{ij}$  is the sample proportion of examinees who answer both items correctly. It can be shown after tedious algebra that

$$r_{ij}^2 = X^2/N, \quad (15.5.10)$$

where  $X^2$  is the quantity commonly used to test for association in a two  $\times$  two

table, and  $N$  is the total number of observations. Thus, in large samples, under the null hypothesis that  $\rho_{ij} = 0$ , the quantity  $Nr_{ij}^2$  is distributed approximately as a chi-square variable with one degree of freedom; or, equivalently, the quantity  $r_{ij}$  is approximately normally distributed with zero mean and variance  $1/N$ . Other statistical properties of  $r_{ij}$  are given by Goodman and Kruskal (1954) and by Berger (1961).

In the special case where  $\pi_i = \pi_j = 0.50$ , we see that

$$\rho_{ij} = 4\pi_{ij} - 1.$$

For a typical pair of items in an aptitude or achievement test,  $\rho_{ij}$  is about 0.10. If the items are of 50% difficulty, it follows that  $\pi_{ij}$  is about 0.275, as illustrated in the two  $\times$  two contingency table, or only barely more than the chance value of 0.25. This table shows how far such typical items are from forming a Guttman perfect scale (see Section 17.3; also Torgerson, 1958, Chapter 12). For a perfect scale, one of the frequencies in the fourfold table would have to be zero.

		$Y_g$
	0	1
$Y_h$	1	0.225 0.275
	0	0.275 0.225
		0.50 0.50

It is important to note that any of the usual formulas for the sample product moment correlation coefficient gives the same numerical results as the one obtained from (15.5.6) and (15.5.9). For binary items, these last equations are the ones usually used to compute such correlations as  $r_{gh}$  and  $r_{gx}$ , which appear in (15.3.2) and (15.3.6). These coefficients may be contrasted with those considered in the following sections.

## 15.6 Biserial Correlation

There are two coefficients, both devised by Karl Pearson, that are widely used both in practical and theoretical work with dichotomous items (although they are seldom used in most other areas of statistics). These are the *biserial* and *tetrachoric* correlation coefficients. Both coefficients are obtained by hypothesizing the existence of a continuous "latent" variable underlying the "right"- "wrong" dichotomy imposed in scoring a dichotomous item. There is some disagreement, even among authorities, on the desirability of hypothesizing such an underlying continuous variable. The reader may judge for himself after reading Chapters 15 and 16, where these coefficients are discussed further.

Consider the bivariate distribution between two variables  $Y'$  and  $\nu$ . The following assumptions are used throughout the section:

**Assumption 15.6.1.**  $Y'$  is normally distributed.

**Assumption 15.6.2.** The regression of  $\nu$  on  $y'$  is linear.

Suppose that  $Y'$  is dichotomized at  $y' = \gamma$ , and suppose that we do not actually observe  $Y'$ ; instead we observe only whether  $Y' < \gamma$  or  $Y' \geq \gamma$ . Assign a

binary variable  $U$  such that  $U = 0$  if  $Y' < \gamma$ , and  $U = 1$  if  $Y' \geq \gamma$ . The problem is to infer  $\rho_{Y'v}$ , the product moment correlation between  $Y'$  and  $v$ , from observations on  $U$  and  $v$ . In item analysis,  $U$  will be the score on a dichotomous item. In this application, the *latent variable*  $Y'$  may be thought of as some measure of the psychological trait or traits that determine success or failure on the item.

If the entire population of values of  $U$  and  $v$  were available, the value of  $\gamma$  could be determined from  $\pi$ , the observed proportion of cases having  $U = 1$ , by using available tables of the normal curve to solve the equation

$$\pi \equiv \bar{\Phi}(\gamma) \equiv \int_{\gamma}^{\infty} \varphi(y') dy', \quad (15.6.1)$$

where  $\bar{\Phi}$  is the normal-curve area defined by (15.6.1), and  $\varphi(t)$  represents the standardized normal density function

$$\varphi(t) \equiv \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2), \quad (15.6.2)$$

$t$  being any mathematical variable.

Given Assumption 15.6.2, the linear regression of  $v$  on  $y'$  must pass through the centroid of each of the two portions into which the line  $y' = \gamma$  divides the bivariate distribution. The centroid of the area on the right of the dichotomy is the point whose coordinates are  $\mathcal{E}(Y' | Y' \geq \gamma)$  and  $\mathcal{E}(v | Y' \geq \gamma)$ ; the centroid on the left is the point whose coordinates are  $\mathcal{E}(Y' | Y' < \gamma)$  and  $\mathcal{E}(v | Y' < \gamma)$ . This is illustrated in Fig. 15.6.1, in which the two centroids are indicated by dots.

If  $Y'$  is expressed in standard-deviation units about its mean, then the mean of a tail of the normal distribution is

$$\begin{aligned} \mathcal{E}(Y' | Y' \geq \gamma) &= \frac{1}{\pi} \int_{\gamma}^{\infty} y' \varphi(y') dy' \\ &= \frac{\varphi(\gamma)}{\pi}, \\ \mathcal{E}(Y' | Y' < \gamma) &= -\frac{\varphi(\gamma)}{1 - \pi}. \end{aligned} \quad (15.6.3)$$

Let the values of

$$\mathcal{E}(v | Y' \geq \gamma) \equiv \mathcal{E}(v | U = 1)$$

and

$$\mathcal{E}(v | Y' < \gamma) \equiv \mathcal{E}(v | U = 0)$$

be denoted by  $\mu^+$  and  $\mu^-$ , respectively. The slope of the regression line of  $v$  on  $y'$

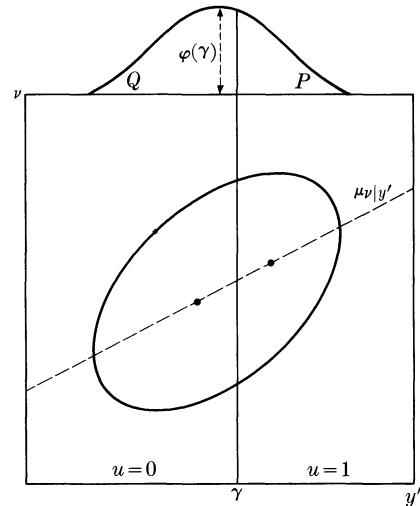


FIG. 15.6.1. Biserial correlation.

passing through the two centroids must then be

$$\beta_{\nu Y'} = \frac{\mu^+ - \mu^-}{\varphi(\gamma)/\pi + \varphi(\gamma)/(1 - \pi)} = \frac{\pi(1 - \pi)}{\varphi(\gamma)} (\mu^+ - \mu^-). \quad (15.6.4)$$

The formula for  $\rho_{Y'\nu}$  can be obtained from that for  $\beta_{\nu Y'}$  by the standard relationship between a correlation and a regression coefficient, that is,  $\beta_{\nu Y'} = \sigma_\nu \rho_{Y'\nu} / \sigma_{Y'}$ . Here  $\sigma_{Y'} = 1$ , because of the scale of measurement chosen in writing (15.6.3). Thus

$$\rho_{Y'\nu} = \frac{\mu^+ - \mu^-}{\sigma_\nu} \frac{\pi(1 - \pi)}{\varphi(\gamma)}. \quad (15.6.5)$$

The coefficient  $\rho_{Y'\nu}$  is the product moment correlation between two continuous variables. For item  $g$ , we shall use the symbol  $\rho'_{gv}$  for the quantity defined by the right-hand side of (15.6.5). The parameter  $\rho'_{gv}$  is called the *biserial correlation coefficient* between the dichotomous variable  $U_g$  and the continuous variable  $\nu$ ; or simply *biserial  $\rho$* . When Assumptions 15.6.1 and 15.6.2 hold, then  $\rho_{Y'\nu} = \rho'_{gv}$ , but this is not the case in general. Biserial  $\rho$  is of interest chiefly when it provides an approximation to  $\rho_{Y'\nu}$ .

We shall denote the sample biserial correlation between  $\nu$  and dichotomous item  $g$  by  $r'_{gv}$ . It is obtained from (15.6.5) by substituting sample values, denoted by  $r'_{gv}$ ,  $M$ ,  $s_\nu$ ,  $h_g$ , and  $p_g$  for population values  $\rho_{Y'\nu}$ ,  $\mu$ ,  $\sigma$ ,  $\gamma$ , and  $\pi$ , respectively. Assuming that  $p_g \neq 0$  or 1, the sample biserial correlation is therefore defined by

$$r'_{gv} \equiv \frac{M^+ - M^-}{s_\nu} \frac{p_g q_g}{\varphi(h_g)}. \quad (15.6.6)$$

The reader should not think of  $r'_{gv}$  as a sample product moment correlation of  $U$  and  $\nu$ , which it is not. He should think of  $r'_{gv}$  as a measure of association between  $U$  and  $\nu$ ; or as a symbol for a certain consistent estimate of  $\rho'_{gv}$ ; or, under Assumptions 15.6.1 and 15.6.2, as an estimate of the product moment correlation between  $\nu$  and  $Y'$ . (Note that  $r'_{gv}$  cannot be an unbiased estimate of  $\rho'_{gv}$  because its value is indeterminate whenever  $p_g = 0$  or  $p_g = 1$ , which will occur with positive probability.)

The present section started with the problem of estimating  $\rho_{Y'\nu}$  from data on  $U$  and  $\nu$ . We obtained an equation valid for population parameters. We then substituted sample statistics for parameters in this equation, and defined the resulting estimate to be the sample biserial  $r$ . This is not necessarily a good way to develop an estimation procedure.

Because of sampling errors, the sample biserial  $r$  may sometimes exceed 1, even when the population sampled satisfies Assumptions 15.6.1 and 15.6.2. This may happen even when  $\rho_{Y'\nu} = \rho'_{gv} = 1$ . For example, to take an extreme case, suppose the sample consists of four observations, the first two both being ( $U = 0, \nu = \nu_0$ ) and the second two both being ( $U = 1, \nu = \nu_1$ ), with  $\nu_1 > \nu_0$ . If these observed values are substituted into (15.6.6),  $\nu_1$  and  $\nu_0$  cancel out and  $r'_{gv} = \frac{1}{2}\varphi(0) \doteq 1.25$ , regardless of the numerical values of  $\nu_1$  and  $\nu_0$ .

A consistent estimator for  $\rho_{Y'v}$ , devised by Brogden and later by Clemans (see Lord, 1963) has the desirable property that it always equals 1 whenever  $\rho_{Y'v} = 1$ . Such an estimate has zero sampling variance when  $\rho_{Y'v} = 1$ . Thus, by comparison, the usual sample biserial  $r$  has *zero efficiency* in this situation; in other words,  $r'_{gv}$  is, by comparison, infinitely inefficient as an estimator. The practical conclusion is that when  $\rho_{Y'v}$  is likely to be high, the Brogden-Clemans estimator is probably preferable to  $r'_{gv}$ . When the correlation in a sample of size  $N$  is positive, the formula for the Brogden-Clemans estimator is

$$\frac{M^+ - M^-}{M_{\max} - M_{\min}}, \quad (15.6.7)$$

where  $M_{\max}$  is the mean of the  $Np$  largest values of  $v$  in the sample and  $M_{\min}$  is the mean of the  $Nq$  smallest values.

Under some further normality assumptions, Tate (1954, 1955a, 1955b) obtained implicit formulas for the maximum likelihood estimator of  $\rho_{Y'v}$  that require data only on  $U$  and  $v$ . These formulas are not simple and will not be given here. Tate also proved that the sample biserial correlation is an asymptotically efficient estimator of  $\rho_{Y'v}$ , when  $\rho_{Y'v} = 0$ ; that is, the sampling variance of  $r'_{gv}$  is asymptotically equal to that of the maximum likelihood estimator when  $\rho_{Y'v} = 0$ . Bhattacharya (1965) has given some further results.

### 15.7 Comparison of Biserial and Point Biserial Coefficients

From (15.6.5) and (15.5.5), we see that the relation between biserial and point biserial correlations is

$$\rho_{gv} = \varphi(\gamma_g)\rho'_{gv}/\sqrt{\pi_g(1 - \pi_g)}. \quad (15.7.1)$$

Thus the point biserial correlation  $\rho_{gv}$  is equal to the biserial correlation multiplied by a factor that depends only on the item difficulty. A parallel equation holds for  $r_{gv}$  and  $r'_{gv}$ .

A survey of normal-curve tables shows that the point biserial is never as much as four-fifths of the biserial. We give a coarse tabulation of the factor  $\varphi(\gamma)/\sqrt{\pi(1 - \pi)}$  as a function of  $\pi$ .

$\pi$	0.50	0.40 or 0.60	0.30 or 0.70	0.20 or 0.80	0.10 or 0.90	0.05 or 0.95
$\frac{\varphi(\gamma)}{\sqrt{\pi(1 - \pi)}}$	0.798	0.79	0.76	0.70	0.58	0.47

This table gives the factor for converting  $\rho_{Y'v}$  to  $\rho_{gv}$  (or  $r'_{gv}$  to  $r_{gv}$ ). Note that since  $\rho_{Y'v} \leq 1$ , the right-hand side of the table also gives the maximum possible value of  $\rho_{gv}$  for any given item difficulty  $\pi$  when the data arise by dichotomizing a population that satisfies Assumptions 15.6.1 and 15.6.2.

Table 15.7.1 presents the data for computing an item-criterion biserial or point biserial correlation for a typical Scholastic Aptitude Test arithmetic reasoning item. The criterion is total test score divided by two. The number of cases is 103,248,  $M^+ = 16.0254$ ,  $M^- = 12.5737$ ,  $p = 0.6502$ ,  $s_p = 5.0901$ ,  $M_{\max} = 17.7131$ , and  $M_{\min} = 9.4361$ . Using (15.6.6), we find the estimated biserial for these data to be  $r'_{gv} = 0.4165$ . The Brogden-Clemans estimator from (15.6.7) is 0.4170.

Both the sample biserial and the sample point biserial correlations between item and test score, or between item and outside criterion, are widely used. Either type of coefficient may be called an important item statistic. The point biserial has the simpler relation to test variance, coefficient  $\alpha$ , and test validity; but (15.7.1) can be used together with (15.3.5), (15.3.8), and (15.4.3) to express test variance, coefficient  $\alpha$ , and validity in terms of biserial correlations.

The point biserial gives the actual product moment correlation between test score, or external criterion, and item. We may view the biserial simply as another measure of association, one different from the product moment correlation. The biserial is widely used because it is hoped that the biserial will demonstrate a type of invariance from one group of examinees to another not provided by the point biserial.

Consider several groups of examinees that differ in level of ability but not in heterogeneity or in other respects. If the necessary assumptions hold, the criterion biserial of a given item will be the same for all groups. But by (15.7.1) the point biserial must be low in any group where the item is very easy or very difficult. Thus the fact that a low point biserial is obtained from one pretest group does not indicate that the item will have low validity for other groups of examinees. Under the assumptions made, however, a low biserial does indicate that the item will have low validity for other groups of examinees.

**Table 15.7.1**  
Relation of item response  
to criterion for an  
arithmetic reasoning item

Criterion	Frequency distribution of examinees responding	
	Incorrectly	Correctly
29	1	130
28	7	393
27	15	705
26	28	1062
25	80	1327
24	124	1729
23	216	2181
22	367	2716
21	585	3262
20	794	3590
19	1174	4139
18	1517	4577
17	1875	4869
16	2224	4928
15	2575	4975
14	2880	4863
13	3143	4375
12	3200	3976
11	3126	3510
10	2981	2887
9	2640	2323
8	2273	1757
7	1735	1277
6	1184	800
5	741	466
4	380	208
3	161	75
2	65	27
1	18	8
0	3	1

The extent to which the assumptions underlying biserial invariance are met in practice cannot be determined from purely mathematical considerations; this is necessarily a matter for empirical investigation. The authors are not aware of definitive empirical studies of this question. However, common experience with the point biserial coefficient confirms the predictions that might plausibly be made from the model illustrated in Fig. 15.6.1 and from Eq. (15.7.1):

1. Items that are very difficult and very easy for a particular group of examinees usually have substantially lower point biserials for that group of examinees than do items of medium difficulty.
2. When items that are very difficult (easy) for one group are given to a slightly more (less) competent group of examinees, the point biserials usually increase.

Figures 15.7.1 and 15.7.2 illustrate the first of these predictions. Figure 15.7.1 shows a fairly typical scatterplot of sample point biserials against item difficulties; Fig. 15.7.2 shows the same data, except that sample biserials are plotted instead of point biserials. The average biserial or point biserial at each difficulty level is shown at the bottom of each figure. The biserials and point biserials shown are with total test score; we should expect similar results for any criterion positively correlated with the items. As is typical in such plots, the biserials for easy items ( $p_g > 0.50$ ) show less tendency to vary with item difficulty than the point biserials.

There is almost always a definite tendency, one that is obvious from Fig. 15.7.2, for difficult multiple-choice items to have low biserials and point bi-

$r_{gx}$	$p_g$	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	Total
		to										
		0.09	0.19	0.29	0.39	0.49	0.59	0.69	0.79	0.89	0.95	
0.90–0.99												
0.80–0.89												
0.70–0.79												
0.60–0.69												
0.50–0.59		/	/	////	//	/						9
0.40–0.49		///	///	///	///	///	///	///	///	///	///	39
0.30–0.39		/	///	///	///	///	///	///	///	///	///	45
0.20–0.29		///	/	///	/		///	///	///	///	///	34
0.10–0.19		///	///	///	/		/	/	/	///	///	22
0.0–0.09		/	/	/	/							6
< 0		/	/	/								4
Total		8	16	13	15	15	15	24	16	24	13	159
Mean $r_{gx}$		0.07	0.16	0.32	0.31	0.43	0.40	0.35	0.33	0.32	0.29	

FIG. 15.7.1. Relation of item-test point biserial correlation to item difficulty for 159 verbal items.

$p_g$	0.05 to 0.09	0.10 to 0.19	0.20 to 0.29	0.30 to 0.39	0.40 to 0.49	0.50 to 0.59	0.60 to 0.69	0.70 to 0.79	0.80 to 0.89	0.90 to 0.95	Total
$r'_{gx}$											
0.90–0.99											
0.80–0.89											
0.70–0.79	/	/						/	/		4
0.60–0.69		///	///	//	///	///	///	///	///	///	23
0.50–0.59	/	///	///	///	///	///	///	///	///	///	41
0.40–0.49	//	/	///	///	///	///	///	///	///	///	35
0.30–0.39	//	///	/	///			///	///	///	//	23
0.20–0.29	//	///	/	/	/	///	///	/	///		19
0.10–0.19	//	/						/			4
0.0–0.09	//	/	/	/							6
< 0	//	/		/							4
Total	8	16	13	15	15	15	24	16	24	13	159
Mean $r'_{gx}$	0.13	0.25	0.44	0.40	0.54	0.50	0.46	0.45	0.48	0.53	

FIG. 15.7.2. Relation of item-test biserial correlation to item difficulty for 159 verbal items.

serials. This is what one would expect with items that can be answered correctly by sheer guessing: If almost all examinees respond to an item more or less at random, the item score cannot correlate very highly with any criterion.

The conclusion reached from practical experience is that biserial correlations tend to be more stable from group to group than point biserials. This is true for any criterion positively correlated with the items. Biserials will still vary from group to group, however, if it is possible to answer the items by guessing. We shall discuss the possibility of "correcting" the biserial for guessing, and of thereby obtaining a more nearly invariant item statistic, in Section 15.11.

A test writer who wishes to build a 149-item verbal test from the 159 items of Figs. 15.7.1 or 15.7.2 and who has no external criterion to guide him might eliminate the ten items with either biserial or point biserial item-test correlations below 0.10; the same items would be eliminated in either case. To build a 93-item test, he might eliminate all items with point biserials below 0.30. Another test writer, using biserials instead of point biserials, might build a 103-item test by eliminating all items with biserials below 0.40. The distribution of item difficulty for the two tests would be

Difficulty	0.05 to 0.09	0.10 to 0.19	0.20 to 0.29	0.30 to 0.39	0.40 to 0.49	0.50 to 0.59	0.60 to 0.69	0.70 to 0.79	0.80 to 0.89	0.90 to 0.95
93-item test	—	1	9	8	14	14	17	11	14	5
103-item test	—	3	9	8	14	14	17	11	16	11

In a crude way (the data are real, not hypothetical), this result illustrates a fact which is obvious from (15.7.1): Use of the point biserial for item selection tends to favor medium-difficulty items over very difficult or very easy items. This may or may not be advantageous: Very difficult and very easy items contribute little to the discriminating power of a mental test, except for examinees near the extremes of the score range.

The sample coefficient  $\alpha$  for the 93-item test and for the 103-item test can be computed from (15.3.8). Of broader interest is the general question: Will selection by item-test point biserials usually produce a more reliable test than selection by biserials? No definite and general answer can be given to this question because test reliability depends not only on item-test correlation but also on item difficulty. Possibly the use of point biserials might tend to produce a more reliable test for groups of examinees exactly like the pretest group, whereas biserials might work better for subsequent groups of examinees that differ somewhat from the pretest group.

Although one commonly prefers items with high biserials, it is theoretically possible for them to be too high. For example, suppose item  $f$  is dichotomous and that  $\rho'_{fv} = 1$ . (Note: Such perfect correlations are not found in cognitive tests.) Then it would be useless to include in the same test a second dichotomous item,  $g$ , of the same difficulty as item  $f$ , and with  $\rho'_{gv} = 1$ . Each examinee will score the same on item  $g$  as on item  $f$ . Thus any item  $h$  with  $0 < \rho'_{hv} < 1$  will be of more use than item  $g$ !

This fact, and some related ones, were pointed out by Tucker (1946). Loevinger (1954) named it the *attenuation paradox*. Specifically the attenuation paradox is the possible decrease in test validity  $\rho_{Xv}$  as a result of an increase in the item validities  $\rho_{gv}$  in certain unusual situations such as the one illustrated above. In Chapter 19 test construction methods by which we can cope with this problem will be outlined; the interested reader may also see Raiffa (1961) and Sitgreaves (1961).

There is no short-cut method that guarantees maximum test reliability. Some methods that have been designed to maximize reliability are given by Thurstone (1931), Myers (1964), and Webster (1953, 1957), among others. A computer program for one method is available from Meyers (1966).

However, maximizing reliability may sometimes be an undesirable goal. For example, a subset of factual items in an achievement test may yield a more reliable score than the total set of items. This can happen, for example, if the other items involve such hard-to-measure but important traits as reasoning ability and creative thinking. Validity is of prime importance in such a case; one would not wish to increase reliability by discarding items if this decreased validity.

A powerful new treatment of the problem of item selection and test construction in the absence of an external criterion is presented in Sections 20.6 and 20.7, for use when the available pool of items meets the necessary assumptions.

### 15.8 Tetrachoric Correlation

Now consider two variables  $Y'_i$  and  $Y'_j$  that are assumed to have a bivariate normal distribution. Suppose that  $Y'_i$  is dichotomized at  $Y'_i = \gamma_i$ , and  $Y'_j$  at  $Y'_j = \gamma_j$ . We do not actually observe values of  $Y'_i$  and  $Y'_j$ , but only whether or not  $Y'_i < \gamma_i$  and  $Y'_j < \gamma_j$ . Let us assign a binary variable  $U_i$  such that  $U_i = 0$  when  $Y'_i < \gamma_i$  and  $U_i = 1$  when  $Y'_i \geq \gamma_i$ ; and let us assign  $U_j$  similarly. The statistical problem is to infer the product moment correlation  $\rho \equiv \rho(Y'_i, Y'_j)$  from observations on  $U_i$  and  $U_j$ . The resulting coefficient is widely used in work with dichotomous test items.

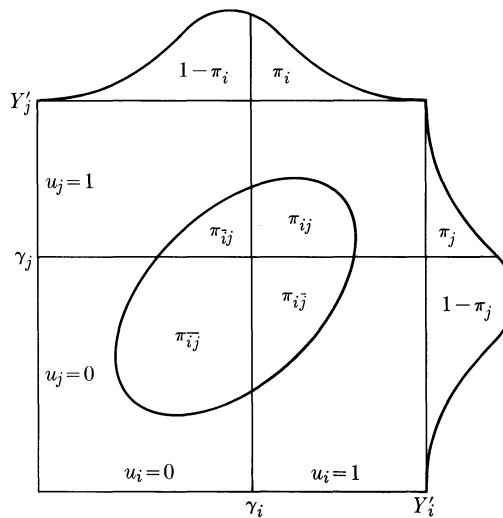


FIG. 15.8.1. Tetrachoric correlation.

Let  $\pi_{ij}$  be the proportion of examinees in the population with  $Y'_i \geq \gamma_i$  and  $Y'_j \geq \gamma_j$ . Let  $\pi_i$  be the proportion with  $Y'_i \geq \gamma_i$ , regardless of  $Y'_j$ , and let  $\pi_j$  be the proportion with  $Y'_j \geq \gamma_j$ , regardless of  $Y'_i$ . Figure 15.8.1 illustrates this situation. Clearly

$$\pi_{ij} \equiv \int_{\gamma_i}^{\infty} \int_{\gamma_j}^{\infty} \varphi(y'_i, y'_j; \rho) dy'_j dy'_i, \quad (15.8.1)$$

$$\pi_i \equiv \int_{\gamma_i}^{\infty} \int_{-\infty}^{\infty} \varphi(y'_i, y'_j; \rho) dy'_j dy'_i, \quad (15.8.2)$$

$$\pi_j \equiv \int_{-\infty}^{\infty} \int_{\gamma_j}^{\infty} \varphi(y'_i, y'_j; \rho) dy'_j dy'_i, \quad (15.8.3)$$

where the integrand represents the standardized bivariate normal distribution

$$\varphi(s, t; \rho) \equiv \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp \left[ -\frac{s^2 + t^2 - 2\rho st}{2(1 - \rho^2)} \right]. \quad (15.8.4)$$

If  $\gamma_i$ ,  $\gamma_j$ , and  $\rho$  are known,  $\pi_{ij}$ ,  $\pi_i$ , and  $\pi_j$  can be determined uniquely from these three equations. Conversely, if  $\pi_{ij}$ ,  $\pi_i$ , and  $\pi_j$  are known, then  $\gamma_i$ ,  $\gamma_j$ , and  $\rho$  are uniquely determined by these equations. When  $\rho$  is computed from  $\pi_i$ ,  $\pi_j$ , and  $\pi_{ij}$  so as to satisfy (15.8.1), (15.8.2), and (15.8.3), it is called the *tetrachoric correlation coefficient*. This coefficient is denoted by  $\rho'_{ij}$ . If  $Y'_i$  and  $Y'_j$  are normal bivariate, then  $\rho'_{ij} = \rho(Y'_i, Y'_j)$ . However,  $\rho'_{ij}$  is defined for any two  $\times$  two table of frequencies, regardless of any underlying  $Y'_i$  and  $Y'_j$ .

If observed sample proportions  $p_{ij}$ ,  $p_i$ , and  $p_j$  are substituted for the  $\pi$  (assuming  $p_i, p_j \neq 0, 1$ ), these formulas determine an estimate of  $\rho_{ij}$  called the *sample tetrachoric correlation coefficient*. This coefficient is denoted by  $r'_{ij}$ .

There are no convenient general formulas for  $\rho'_{ij}$  or  $r'_{ij}$ . One computational method is to invert the infinite series (15.9.2) given below. Usually one has recourse to published tables, such as that of the United States National Bureau of Standards (1959), from which  $\rho'_{ij}$  or  $r'_{ij}$  can be approximated; or to electronic computer programs. Tallis (1962) has derived a maximum likelihood estimate of  $\rho_{ij}$  but we shall not treat this estimate here. Other methods for estimating  $\rho(Y'_i, Y'_j)$  may be found in Ross and Weitzman (1964). For the sampling error of  $r'_{ij}$ , see Hayes (1943).

### 15.9 A Comparison of Tetrachoric and Phi Coefficients

Let  $\gamma_i$  and  $\gamma_j$  be determined from  $\pi_i$  and  $\pi_j$  by the relation

$$\pi_g \equiv \int_{\gamma_g}^{\infty} \varphi(y') dy'. \quad (15.9.1)$$

The following infinite series (see Kendall and Stuart, 1961, Eq. 26.67), developed by Karl Pearson (1900), relates the phi coefficient  $\rho_{ij}$  and the tetrachoric correlation  $\rho'_{ij}$ :

$$\begin{aligned} \frac{\sigma_i \sigma_j \rho_{ij}}{\varphi(\gamma_i) \varphi(\gamma_j)} &= \rho'_{ij} + \frac{1}{2} \gamma_i \gamma_j \rho'_{ij}^2 + \frac{1}{6} (\gamma_i^2 - 1)(\gamma_j^2 - 1) \rho'_{ij}^3 \\ &\quad + \frac{1}{24} \gamma_i \gamma_j (\gamma_i^2 - 3)(\gamma_j^2 - 3) \rho'_{ij}^4 \\ &\quad + \frac{1}{120} (\gamma_i^4 - 6\gamma_i^2 + 3)(\gamma_j^4 - 6\gamma_j^2 + 3) \rho'_{ij}^5 + \dots \end{aligned} \quad (15.9.2)$$

In the special case where  $\pi_i = \pi_j = 0.50$ , it can be shown that

$$\rho'_{ij} = \sin\left(\frac{\pi \rho_{ij}}{2}\right), \quad (15.9.3)$$

where  $\pi \doteq 3.1416$  and the quantity  $\pi \rho_{ij}/2$  is expressed in radians. Table 15.9.1 shows the relation between  $\rho'_{ij}$  and  $\rho_{ij}$  when  $\pi_i = \pi_j = 0.50$ . Intercorrelations between aptitude- or achievement-test items are seldom larger than  $\rho'_{ij} = 0.50$ . An approximation to  $\rho'_{ij}$  is

$$\rho'_{ij} \doteq 3\rho_{ij}/2; \quad (15.9.4)$$

the approximation is close when  $\pi_i = \pi_j = 0.50$  and  $\rho'_{ij} \leq 0.50$ .

**Table 15.9.1**

Corresponding values for two measures of association between dichotomous items of 50% difficulty

Tetrachoric correlation $\rho'_{ij}$	0.10	0.20	0.30	0.50	0.80	0.96	1.00
Phi coefficient $\rho_{ij}$	0.06	0.13	0.19	0.33	0.59	0.82	1.00

Let

$$\begin{cases} \pi_{ij} \equiv \text{Prob } (u_i = 1, u_j = 0), \\ \pi_{\bar{i}j} \equiv \text{Prob } (u_i = 0, u_j = 1), \\ \pi_{\bar{i}\bar{j}} \equiv \text{Prob } (u_i = 0, u_j = 0), \end{cases} \quad (15.9.5)$$

and let  $p_{ij}$ ,  $p_{\bar{i}j}$ , and  $p_{\bar{i}\bar{j}}$  be the corresponding observed proportions. The familiar facts about product moment correlations make it clear that  $\rho_{ij} \equiv \rho(u_i, u_j) = 1$  if and only if  $\pi_{ij} = \pi_{\bar{i}j} = 0$ . It follows that two dichotomous items of unequal difficulty ( $\pi_i \neq \pi_j$ ) cannot have  $\rho_{ij} = 1$ . This is as it should be, since  $\pi_i \neq \pi_j$  implies that either  $\pi_{ij} \neq 0$  or  $\pi_{\bar{i}j} \neq 0$ . In fact it is easily shown that when  $\pi_i \geq \pi_j$ , the maximum value of  $\rho_{ij}$ , denoted by  $\varphi_{\max}$ , is

$$\varphi_{\max} \equiv \sqrt{\pi_j(1 - \pi_i)/\pi_i(1 - \pi_j)}. \quad (15.9.6)$$

When will  $\rho'_{ij} = \pm 1$ ? If we consider Fig. 15.8.1, we know that as  $\rho(Y'_i, Y'_j)$  becomes larger and larger in absolute value, the ellipse representing the contours of the bivariate normal distribution becomes longer and thinner. When  $\rho(Y'_i, Y'_j)$  becomes  $\pm 1$ , the ellipse becomes a straight line on which all the frequency lies. It is geometrically clear that this line, and the frequencies on it, can fall in at most three of the four quadrants produced by the double dichotomization. Thus at least one quadrant will have zero frequency whenever  $\rho(Y'_i, Y'_j) = \pm 1$ .

Conversely, a knowledge of the properties of the bivariate normal distribution [ruling out the degenerate case where  $\sigma(Y'_i)$  or  $\sigma(Y'_j) = 0$ ] makes it clear that there will always be some frequency in each of the four quadrants as long as  $\rho(Y'_i, Y'_j) \neq \pm 1$ . Thus a necessary and sufficient condition for  $\rho'_{ij} = \pm 1$  is that at least one of the four values  $\pi_{ij}$ ,  $\pi_{\bar{i}j}$ ,  $\pi_{ij}$ , and  $\pi_{\bar{i}\bar{j}}$  shall be zero.

Both the phi coefficient and the tetrachoric correlation are widely used for dichotomous test items. The phi coefficient has a simple relation to item-test correlation (15.3.7), to test variance (15.3.1), and through these to test reliability. In the special case where all test items provide parallel measurements, the Spearman-Brown formula (4.4.9) holds and the test reliability  $\rho_{XX'}$  is a very simple function of the phi coefficient  $\rho_{ij}$ :

$$\rho_{XX'} = \frac{n\rho_{ij}}{1 + (n - 1)\rho_{ij}}. \quad (15.9.7)$$

Because of (15.9.2) and (15.7.1), the test variance, reliability, and item-test biserials can also be considered functions of the tetrachoric item intercorrelations, but not in any simple way.

The objection to tetrachorics is the assumption of underlying bivariate normality, which clearly cannot be expected to hold with any generality. The fact is that most groups of examinees, particularly at the college level, have been subjected to selection, frequently drastic selection. Even if the assumption of bivariate normality were to hold for one group of examinees for all pairs of items, it could hardly be expected to hold for another group selected by a different method.

Tetrachorics are widely used because it is hoped that they will show, approximately, certain invariance properties not provided by phi coefficients. Although the invariance of tetrachorics in actual practice is a matter for empirical investigation, some practical conclusions about the invariance properties of phi coefficients can be deduced from mathematical considerations. The tetrachoric correlation is designed to remain invariant for groups of examinees that have different average ability levels on variables  $Y'_i$  and  $Y'_j$ , but that otherwise have the same bivariate normal distribution for these two variables. Since the assumption of bivariate normality is gratuitous, it is appropriate to ask: Is there some bivariate frequency distribution that gives to the phi coefficient the same desirable invariance properties that the bivariate normal gives to the tetrachoric? A theorem of Carroll's (1961) answers this question.

**Theorem 15.9.1.** *There is no bivariate frequency distribution that, when doubly dichotomized, yields the same phi coefficient regardless of the points of dichotomy.*

*Proof.* Suppose that  $\varphi_0$  is the phi coefficient for some dichotomization. It is clear from (15.9.6) that we can always find a  $\pi_j$  such that  $\varphi_{\max} < \varphi_0$ . It follows that the dichotomization producing  $\pi_j$  must yield  $\rho_{ij} < \varphi_0$ . This proves the theorem.  $\square$  (The comparable theorem for point-biserial coefficients is false.)

To meet this situation, some workers use the ratio  $\rho_{ij}/\varphi_{\max}$  as a measure of correlation. Is there a bivariate frequency distribution  $f(y'_i, y'_j)$  that allows this ratio to remain invariant regardless of the dichotomization? Carroll (1961) has shown that there exist  $f(y'_i, y'_j)$  satisfying this requirement. It can further be shown that for any given  $\rho'_{ij}$  and for any bounded discrete marginal distributions  $f(y'_i) = f(y'_j)$ , the desired  $f(y'_i, y'_j)$  exists and is uniquely determined. As Carroll points out, however, when  $\rho'_{ij} \neq 0$ , these  $f(y'_i, y'_j)$  are of such peculiar shape as to be completely implausible. The reader may visualize a typical shape with the aid of Fig. 15.9.1 if he will think of the frequency in each cell as concentrated at a point rather than distributed evenly over the whole cell. This bivariate distribution has the surprising but probably useless property that its product moment correlation remains invariant under any monotonic transformation of  $Y'_i$  and  $Y'_j$ , provided that the same transformation is used for both variables.

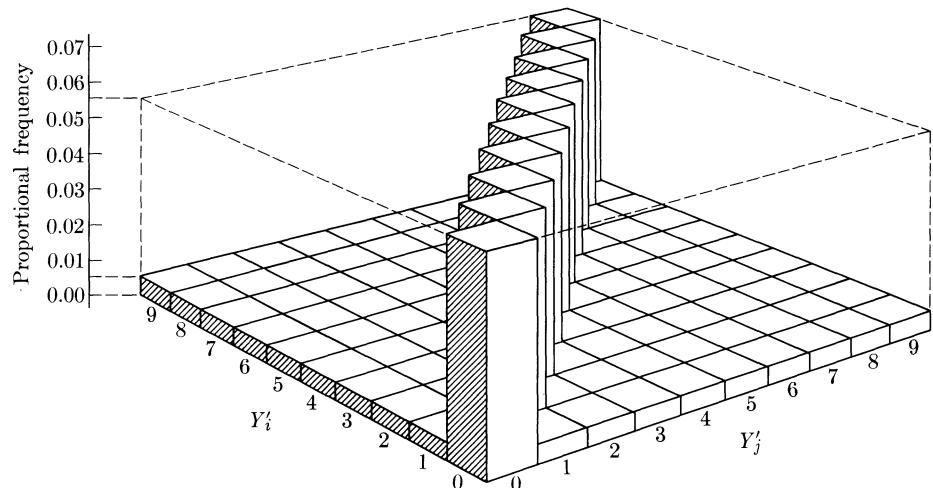


FIG. 15.9.1. A distribution that yields the same  $\rho_{ij}/\varphi_{\max}$  wherever dichotomized. [From J. B. Carroll, The nature of data, or how to choose a correlation coefficient. *Psychometrika*, 1961, **26**, 347-372. Used by permission.]

The main conclusion is that neither the phi coefficient nor  $\rho_{ij}/\varphi_{\max}$  has desirable invariance properties. On the other hand, tetrachorics sometimes have these properties, but only for certain populations of examinees where the necessary assumptions are satisfied.

Except for factor analyses of test items, neither phi coefficients nor tetrachorics are widely used in practical item analysis work, since both describe a pair of items rather than a single item and since  $n(n - 1)/2$  coefficients are needed to describe an  $n$ -item test. Both coefficients are of theoretical importance and will be used in subsequent chapters.

If the matrix of phi coefficients for a set of items should happen to be of rank one for some particular group of examinees, then its rank is likely to be much larger than one for any other group in which the item difficulties are altered. For this reason, phi coefficients cannot be recommended for factor analysis. The same objection applies to the tetrachoric coefficients, however, unless the necessary assumptions relating to multivariate normality are satisfied. For relevant empirical results, see Guilford (1941), Wherry and Gaylord (1944), Gourlay (1951), and Carroll (1961).

If  $Y'_i$  and  $Y'_j$  are bivariate normal for all  $i, j$ , then clearly the population matrix  $\|\rho_{ij}\| = \|\rho(Y'_i, Y'_j)\|$  will be Gramian, that is, nonnegative definite. Although a matrix of sample phi coefficients is always Gramian, a matrix of sample tetrachorics is often non-Gramian (even when the population matrix is Gramian). Difficulties may arise if certain common statistical techniques are incautiously applied to non-Gramian sample tetrachoric matrices. However, this is an empirical difficulty with the statistical technique used rather than a theoretical objection to the use of tetrachoric correlations.

### 15.10 Considerations in the Choice of Test Construction Techniques

In previous sections, we have discussed the problem of item selection in situations where an external criterion is and is not available. The methods presented in those sections involve no item selection algorithm, however, though we mentioned some item indexes that test constructors have found useful in selecting items for a test. It seems fair to say that in much test construction work these indexes receive only moderate attention and that the ultimate decision to include or exclude a given test item depends in large part on the subjective judgment of the test constructor; it is he who decides whether or not the item will contribute to prediction, or whether it measures what he wishes to measure. In the case of achievement tests, particularly, the selection and validation of items is based largely on a judgment of how adequately the set of items spans the domain of interest. The test constructor then uses item statistics primarily to eliminate bad items or to choose the best item from a group of items which he presumes to measure some particular aspect of the domain of interest.

There are, in the extreme, two quite different approaches to test construction. The first of these is essentially empirical. Items are selected for inclusion solely on the basis of their contribution to the overall empirical validity of the test with respect to some specified criterion. Considerations of item validity and, to a lesser extent, item intercorrelation are paramount; specific items are selected because they have high correlations with the criterion and, to the extent possible, low intercorrelations, just as tests are selected for a battery. Most commonly, for reasons of economy, multiple regression methods are not used to determine weights for the individual items; rather, unit weights and approximate methods are employed, as described in Section 15.4. If an examiner desires optimal weights, the regression methods for selecting and weighting tests for a battery that were presented in Chapters 12 and 13 may be used to select and weight items for a test. Section 20.8 provides a general answer in one context to the question of how much is gained by using an optimally weighted composite of item scores rather than an unweighted composite.

The use of weighted item scores should be particularly profitable in any situation where the test is composed of multiple-choice items, so that there is some positive chance level of success, and where the examiner's purpose is to discriminate as effectively as possible among examinees who score near the chance level. The performance of less able examinees on the difficult items is then almost entirely a matter of random guessing. Since scoring such items simply adds noise to the total test score, a more valid score for such examinees will be obtained by assigning zero or near-zero weights to the difficult items.

Of course, when such an empirically determined test is constructed, it ought to be validated as a whole. As in selection of tests by multiple regression methods, the cross validation of the final predictor typically displays substantial shrinkage in comparison with the same predictor in the original sample because the items have been selected in part on the basis of sampling fluctuations (see Section 13.2). The empirical validities of such tests may be relatively high, but to the

extent that this empirical scaling method selects items on the basis of *low* item intercorrelation, such tests clearly cannot be supposed to measure any *one* underlying psychological trait. Hence such tests find little application in theoretical psychological research, though they often prove very useful in the solution of technological problems.

The second and opposite approach to test construction is less purely statistical and more theoretical-psychological. This approach uses psychological theory to suggest certain unitary traits that are important to the study of human behavior. Test scales are then constructed to measure these unitary traits, and there is a corresponding emphasis on selecting items that "measure the same thing". Here biserial and tetrachoric item-test correlations, as described in Sections 15.6 and 15.8, assume a greater value: When several scales are being constructed simultaneously, an item may be assigned to a specified scale only if it correlates highly with the items of that scale and does *not* correlate highly with the items of the other scales. However, in this approach it is frequently true that no specific criterion is used in test construction and hence no item validities are available. In these cases, very heavy reliance is placed on the psychological skill of the test constructor. Questions of validity must then be carefully investigated *after* the scale has been constructed.

Actually most test construction projects involve some mixture of these two approaches. For example, in aptitude testing some reliance is typically placed on the judgment of experts to delineate the specific topics covered by the test and often to construct preliminary forms of the questions. These items may then be subject to editorial revision to cast them in proper form. These items are then pretested, sometimes with, but often without also obtaining, a criterion measure, and the final items for the test are selected largely on the basis of item-test correlations. Where a criterion measure is not actually obtained, there is a presumption, based on previous validity studies, that these experts will be able to produce valid items. When such tests are constructed year after year, each such form is usually not subjected to detailed empirical validation. Rather, if it can be assumed that the actual test construction procedure and the kind of examinees being measured do not change too much on a yearly basis, then such validity is typically demonstrated only periodically, or validity may be inferred indirectly by correlating the new form with an old form. Thus the process of test construction is typically a mixture of subjective and empirical approaches, with validation being provided often enough to insure the quality of the test.

We shall not pursue here the long-standing controversy between the advocates of empirical scaling methods and the advocates of homogeneous scaling methods. Homogeneous scales are very desirable for many purposes. Their advantages in theory construction and theoretical investigations seem to be accepted by most psychologists interested in such questions. We shall treat the mathematical and statistical advantages of isolating pools of relatively homogeneous items in Chapters 16 through 20.

This is not to say that questions of validity are foreign to the construction and evaluation of homogeneous scales. Indeed, since such scales typically have little empirical basis for their construction, questions of empirical validity are particularly important for these scales. Furthermore the fact that a scale is homogeneous and that its items appear, on the face of the matter, to be concerned with the trait in question does not establish the worth of the scale. In fact the appropriateness of such scales can be established only by a series of validity (and reliability) studies that show that

- 1) the scale correlates well with those empirical variables that are presumed to depend on the trait in question,
- 2) the scale correlates with other scales with which theory suggests it should correlate, and
- 3) the scale does not correlate with scales with which theory suggests it should not correlate.

The first of these validities is an *empirical validity*. The second and third validities are called *convergent validity* and *discriminant validity* by Campbell and Fiske (1959), who presented detailed specifications for study of these two types of validity. Since they are topics in the theory of construct validity, these methods are discussed in more detail in Chapter 12.

Writers such as Messick (1964) argue convincingly that "empirical validity is not enough", particularly where personality measures are concerned. On the other hand, even if empirical validity is not considered to be a sufficient criterion to establish the goodness of a test, we do believe that empirical validity, of some sort, is a necessary criterion. As we have indicated in Chapter 2, all theoretical constructs must either have a direct behavioral reference or be linked theoretically with another construct that has such a reference. Thus we would require that every test have a direct, or indirect, empirical validity. Whether a test is constructed to be marketed as a theoretical or technological tool or as an instrument for some *ad hoc* research or prediction purpose, the relevant validation requirements of the Standards for Educational and Psychological Tests and Manuals should not be ignored. Rather, as Dobbin (1966) has suggested, these standards should also be interpreted "as standards for the users of tests".

### 15.11 Formula Scoring and Corrections for Chance Success

If  $NP$  examinees know the answer to an item and actually answer it correctly, and if the remaining  $NQ$  guess at random with chance of success  $1/A$  for each examinee, then an unbiased estimate of the proportion  $P$  is

$$\hat{P} = p - \frac{q}{A - 1}, \quad (15.11.1)$$

where  $p$  is the sample item difficulty (proportion of right answers) and  $q \equiv 1 - p$ .

Some writers have recommended the use of  $\hat{P}$  (or a statistic derived from  $\hat{P}$ ) as an index of item difficulty corrected for chance success.

Unfortunately  $\hat{P}$  not only may be, but sometimes actually is, negative. This may happen because of sampling fluctuations, or because certain examinees do not guess at random but systematically choose a plausible wrong answer in preference to the right one. In the latter case (which is not unusual), the assumptions underlying the formula for  $\hat{P}$  are clearly violated. In view of all this, the correction for chance success cannot be recommended for routine use.

*Items that can be answered correctly by random guessing do not fit the biserial correlation model.* No matter how low he is on the criterion variable ( $v$  or  $X$ ), the examinee always has a chance of at least  $1/A$  to answer the item correctly. This situation is not compatible with Assumption 15.6.2, which states that the criterion variable has a linear regression on a variable ( $Y'$ ) with infinite range.

Formulas for correcting the data for chance success (see Plumlee, 1952) attempt to meet this objection to the use of biserial correlation. The point biserial can be corrected in the same way. Carroll (1945) has given formulas for correcting two  $\times$  two contingency tables for random guessing. The objections to  $\hat{P}$  raised above, however, apply to any correction of the biserial, the point biserial, the tetrachoric, and the phi coefficient. For further discussion, the interested reader should refer to Plumlee's empirical works (1952, 1954), which illustrate the theoretical and practical uses of these corrections and the difficulties encountered in their application.

The foregoing discussion holds only for dichotomous items. Suppose a test is scored by a formula such as (14.3.4):

$$\hat{k}_a \equiv x_a - \frac{w_a}{A - 1},$$

where  $x_a$  is the number of right answers for examinee  $a$ ,  $w_a$  is the number of wrong answers, and  $\hat{k}_a$  is the assigned test score. This formula assigns a score of +1 to right answers, a score of 0 to omitted answers, and a score of  $-1/(A - 1)$  to wrong answers. Clearly *such a formula score on a test is not a sum of dichotomous item scores. Consequently none of the correlation coefficients discussed above are applicable.* Those formulas in the first four sections of this chapter that are not specialized to dichotomous test items can still be applied, but these formulas contain as terms general product moment correlations rather than point biserials and phi coefficients. Formulas for "triserial" (see Jaspen, 1965, for references) and "polychoric" (Pearson, 1900) correlation coefficients, generalizing the biserial and tetrachoric coefficients, are available in the literature but are seldom used in item analysis.

## 15.12 Invariant Item Parameters

The item parameters treated in this chapter will no doubt continue to play an important part in any study of the relation of test characteristics to item characteristics. On the other hand, we seem to need some better approach to the problem of securing item parameters that might remain invariant in a variety of

practical situations. A consideration of such parameters is an incidental outcome of the theory developed in succeeding chapters.

It is clear that the mean score (or expected score) on item  $g$  cannot be expected to remain invariant when the population of examinees is substantially changed. If two groups of examinees are at different levels with respect to the abilities or traits involved in answering an item, then the mean score or difficulty of the item will be different for the two groups.

Similar statements can be made about the other item parameters already mentioned: item variance, item-test correlation, and item validity. If one group of examinees is more homogeneous than another on the traits involved in answering an item, the variance is likely to be smaller for the more homogeneous group. The same is true of item-test correlation and item validity, because of the well-known effect of homogeneity on correlation coefficients (see Chapter 6).

These facts create practical problems in utilizing item-analysis results, since item parameters obtained from several discrepant pretest groups of examinees must in practice be used in setting up tests to be administered to still other groups of examinees. What are needed are item parameters that remain approximately invariant from group to group.

Since this need arises because of variations among groups of examinees in the abilities or traits measured by the items, any solution must necessarily involve a consideration of the relation between these abilities or traits and examinee performance on the items. The problem of dealing with the relationship between the examinee's mental traits and his performance is not a simple one, but we cannot avoid it. It lies at the heart of mental test theory, which is, after all, fundamentally concerned with inferring the examinee's mental traits from his responses to test items. A treatment of this basic task of inference is the chief concern of Chapters 16 through 20.

These chapters, incidentally, will define a set of item parameters with the desired invariance properties. Methods currently available for estimating such parameters in practical work leave something to be desired (see Section 17.9). Practical workers will in many cases continue to work with the item parameters discussed in the present chapter, using the theories outlined in Chapters 16 to 20 to judge the degree and type of deviations from invariance to be expected in their practical applications.

## References and Selected Readings

- BERGER, AGNES, On comparing intensities of association between two binary characteristics in two different populations. *Journal of the American Statistical Association*, 1961, **56**, 889-908.
- BHATTACHARYA, P. K., Asymptotic efficiencies of some tests of independence used in item selection. In H. Solomon (Ed.), *Item analysis, test design, and classification*. U.S. Office of Education, Cooperative Research Program, Project No. 1327. Stanford: Stanford University Press, 1965, pp. 98-136.

- CAMPBELL, D. T., and D. W. FISKE, Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81-105.
- CARROLL, J. B., The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 1945, **10**, 1-19.
- CARROLL, J. B., The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 1961, **26**, 347-372.
- CURETON, E. E., Corrected item-test correlations. *Psychometrika*, 1966, **31**, 93-96.
- DARLINGTON, R. B., and C. H. BISHOP, Increasing test validity by considering inter-item correlations. *Journal of Applied Psychology*, 1966, **50**, 322-330.
- DAS GUPTA, S., Point biserial correlation coefficient and its generalization. *Psychometrika*, 1960, **25**, 393-408.
- DOBBIN, J. E., Review of standards for educational and psychological tests and manuals. *Educational and Psychological Measurement*, 1966, **26**, 751-753.
- ELFVING, G., The item-selection problem and experimental design. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961, pp. 81-87. (a)
- ELFVING, G., Item selection and choice of nonrepeatable observations for estimation. *Op. cit.*, pp. 88-95. (b)
- ELFVING, G., Contributions to a technique for item selection. *Op. cit.*, pp. 96-108. (c)
- ELFVING, G., ROSEDITH SITGREAVES, and H. SOLOMON, Item-selection procedures for item variables with a known factor structure. *Psychometrika*, 1959, **24**, 189-205.
- ELFVING, G., ROSEDITH SITGREAVES, and H. SOLOMON, Item-selection procedures for item variables with a known factor structure. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961, pp. 64-80.
- GOODMAN, L. A., and W. H. KRUSKAL, Measures of association for cross classifications. *Journal of the American Statistical Association*, 1954, **49**, 732-764.
- GOURLAY, N., Difficulty factors arising from the use of tetrachoric correlations in factor analysis. *British Journal of Psychology*, Statistical Section, 1951, **4**, 65-76.
- GREEN, B. F., JR., A note on item selection for maximum validity. *Educational and Psychological Measurement*, 1954, **14**, 161-164.
- GUILFORD, J. P., The difficulty of a test and its factor composition. *Psychometrika*, 1941, **6**, 67-77.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.
- GUTTMAN, L., The quantification of a class of attributes: a theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment*. Social Science Research Council, Bulletin 48, 1941, pp. 321-345.
- HAYES, S. P., JR., Tables of the standard error of tetrachoric correlation coefficient. *Psychometrika*, 1943, **8**, 193-203.
- HENRYSSON, S., Correction of item-total correlations in item analysis. *Psychometrika*, 1963, **28**, 211-218.
- HORST, A. P., Item analysis by the method of successive residuals. *Journal of Experimental Education*, 1934, **2**, 254-263.

- HORST, P., Determination of optimal test length to maximize the multiple correlation. *Psychometrika*, 1949, **14**, 79-88.
- HORST, P., *Psychological measurement and prediction*. Belmont, Calif.: Wadsworth, 1966.
- HORST, P., and CHARLOTTE MACEWAN, Predictor elimination techniques for determining multiple prediction batteries. *Psychological Reports*, 1960, **7**, 19-50.
- JASPERN, N., Polyserial correlation programs in FORTRAN. *Educational and Psychological Measurement*, 1965, **25**, 229-233.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. Vol. 2: *Inference and relationship*. New York: Hafner, 1961.
- LEV, J., Maximizing test battery prediction when the weights are required to be non-negative. *Psychometrika*, 1956, **21**, 245-252.
- LINHART, H., A criterion for selecting variables in a regression analysis. *Psychometrika*, 1960, **25**, 45-58.
- LOEVINGER, JANE, The attenuation paradox in test theory. *Psychological Bulletin*, 1954, **51**, 493-504.
- LORD, F. M., Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, 1958, **23**, 291-296.
- LORD, F. M., Biserial estimates of correlation. *Psychometrika*, 1963, **28**, 81-85.
- MESSICK, S., Personality measurement and college performance. In *Proceedings of the 1963 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1964. Reprinted in Anne Anastasi, (Ed.), *Testing problems in perspective*. Washington, D.C.: The American Council on Education, 1966.
- MEYERS, E. D., JR., IBM 1401 stepwise scaling programs. *Computers in Behavioral Science*, 1966, **11**, 319-320.
- MYERS, C. T., Item analysis and test reliability. *Research Bulletin 64-8*. Princeton, N.J.: Educational Testing Service, 1964.
- PEARSON, K., On the correlation of characters not quantitatively measurable. *Royal Society Philosophical Transactions*, Series A, 1900, **195**, 1-47.
- PLUMLEE, LYNNETTE B., The effect of difficulty and chance success on item-test correlation and on test reliability. *Psychometrika*, 1952, **17**, 69-86.
- PLUMLEE, LYNNETTE B., The predicted and observed effect of chance success on multiple-choice test validity. *Psychometrika*, 1954, **19**, 65-70.
- RAIFFA, H., Statistical decision theory approach to item selection for dichotomous test and criterion variables. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961, pp. 187-220.
- RICHARDSON, M. W., Relation between the difficulty and the differential validity of a test. *Psychometrika*, 1936, **1**, 33-49.
- ROSS, J., and R. A. WEITZMAN, The twenty-seven percent rule. *Annals of Mathematical Statistics*, 1964, **35**, 214-221.
- ROZEBOOM, W. W., *Foundations of the theory of prediction*. Homewood, Ill.: Dorsey, 1966.
- SAUPE, J. L., Selecting items to measure change. *Journal of Educational Measurement*, 1966, **3**, 223-228.

- SITGREAVES, ROSEDITH, A statistical formulation of the attenuation paradox in test theory. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961, pp. 17-28.
- STOUFFER, S. A. (Ed.), Measurement and prediction. *Studies in social psychology in World War II*, Vol. IV. Princeton, N.J.: Princeton University Press, 1950.
- SUMMERFIELD, A., and A. LUBIN, A square root method of selecting a minimum set of variables in multiple regression. *Psychometrika*, 1951, **16**, 271-284.
- TALLIS, G. M., The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 1962, **18**, 342-353.
- TATE, R. F., Correlation between a discrete and a continuous variable. Point biserial correlation. *Annals of Mathematical Statistics*, 1954, **25**, 603-607.
- TATE, R. F., Application of correlation models for biserial data. *Journal of the American Statistical Association*, 1955, **50**, 1078-1095. (a)
- TATE, R. F., The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, 1955, **42**, 205-216. (b)
- THURSTONE, L. L., *The reliability and validity of tests*. Ann Arbor, Mich.: Edwards Brothers, 1931.
- TORGERSON, W. S., *Theory and methods of scaling*. New York: Wiley, 1958.
- TUCKER, L. R., Maximum validity of a test with equivalent items. *Psychometrika*, 1946, **11**, 1-13.
- United States National Bureau of Standards, Tables of the bivariate normal distribution function and related functions. Applied mathematics series 50. Washington, D.C.: United States Government Printing Office, Division of Public Documents, 1959.
- WEBSTER, H., Approximating maximum test validity by a non-parametric method. *Psychometrika*, 1953, **18**, 207-211.
- WEBSTER, H., Maximizing test validity by item selection. *Psychometrika*, 1956, **21**, 153-164.
- WEBSTER, H., Item selection methods for increasing test homogeneity. *Psychometrika*, 1957, **22**, 395-403.
- WHERRY, R. J., and R. H. GAYLORD, Factor pattern of test items and tests as a function of the correlation coefficient: content difficulty and constant error factors. *Psychometrika*, 1944, **9**, 237-244.
- WOLF, R., Evaluation of several formulae for correction of item-total correlations in item analysis. *Journal of Educational Measurement*, 1967, **4**, 21-26.

## CHAPTER 16

# LATENT TRAITS AND ITEM CHARACTERISTIC FUNCTIONS

### 16.1 Introduction

Any theory of latent traits supposes that an individual's behavior can be accounted for, to a substantial degree, by defining certain human characteristics called *traits*, quantitatively estimating the individual's standing on each of these traits, and then using the numerical values obtained to predict or explain performance in relevant situations. For example, to predict a person's performance in a graduate program in psychometrics, we would want to know his scores on a particular set of traits. The traits of interest would include quantitative aptitude, verbal aptitude, mathematical ability, knowledge of the subject matter of psychology, and also, perhaps, the less tangible traits that we call perseverance and creativity. Traits such as finger dexterity and musical aptitude would be of no interest.

Much of psychological theory is based on a trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behavior.

The problem of identifying and defining traits in terms of observable variables and determining which are important in a given behavioral context is a major problem in any theory of psychology. In studying any particular behavioral context, the psychologist will bring to bear the relevant theoretical formulation, and will perhaps employ one or more models belonging to that theory in designing experimental studies and analyzing the data from these studies; it is such interplay of theory, model, and data that adds to scientific knowledge.

The psychologist who wishes to use an examinee's responses to a mental test to make inferences about his psychological traits must have some knowledge of how psychological traits determine or are related to examinees' responses. The purpose of the present chapter is to introduce a type of mathematical model that is used to describe the relationship between an examinee's psychological traits and his responses.

A very general treatment is given in the first three sections of this chapter. The next five sections are devoted to describing the normal ogive model, which

is a specialization of the general model. We present a sufficient condition for the normal ogive model to give the reader some confidence in the plausibility of the model. The subject of the next two sections is the relation of the invariant item parameters of the model to the item parameters discussed in the preceding chapter.

The last four sections of this chapter describe how the relation between psychological trait and test score depends on the characteristics of the items put into the test. These sections are not restricted to the normal ogive model. We believe this type of analysis to be basic to any consideration of a psychological test as a measuring instrument.

In conventional testing, the same test is administered to many examinees. In the tests of the future, it is possible that each examinee will take a different set of items, selected to measure his characteristics as efficiently as possible. For such testing, some of the test theory of Chapters 1 through 13 and 21 through 23 will be inadequate. Chapters 16 through 20 are for the most part equally applicable to conventional testing and to situations where different examinees take different sets of test items.

## 16.2 Latent Variables

Consider a set of  $n$  items and a set of  $k'$  traits. We assume that each of the traits affects examinee performance on at least one item in the set. These traits are to be thought of as psychological dimensions necessary for the psychological description of individuals.

These traits are called *latent traits* or *latent variables*. We shall denote them by the vector

$$\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \dots, \theta_{k'}).$$

We can now represent each examinee by a point in a  $k'$ -dimensional space, called a *latent space*. We reserve the question of what metric to use for locating such points for later discussion.

Next, consider all the examinee populations that may be of interest for this set of  $n$  items. Assume that each item is administered just once to each examinee, and consider the conditional frequency distribution (over people) of item score  $Y_{g*}$  for any fixed value of  $\boldsymbol{\theta}$ . If this (unobservable) distribution is not the same for all these populations of examinees, then there are one or more psychological dimensions in addition to  $\theta_1, \theta_2, \dots, \theta_{k'}$  that discriminate among the populations of interest. In defining the *complete latent space*, therefore, we must include these additional dimensions. *Thus, by definition, in the complete latent space the conditional distribution of item score for fixed  $\boldsymbol{\theta}$  is the same for all the populations of interest.* We shall denote the dimensionality of the complete latent space by  $k$ .

The reader will have noted that the nature and dimensionality of the complete latent space have been defined in such a way that they depend both on

the particular traits and on the particular populations of interest to the psychometrician. The psychometrician is likely to wish to define his complete latent space to include all "important" psychological dimensions that affect performance on the given set of items and to exclude those variables that comprise "errors of measurement". Unfortunately it seems to be logically impossible to distinguish objectively those variables that are simply "errors of measurement" from those that are not. For the purposes of this chapter, then, we shall define the complete latent space to include all those variables that discriminate among "relevant" groups of examinees, regardless of the psychometrician's subjective definition of "errors of measurement". Hereafter, the vector  $\boldsymbol{\theta}$  will include  $k$  elements and will refer to the complete latent space.

The regression of item score on  $\boldsymbol{\theta}$  is called the *item characteristic function*. Because of its definition, the item characteristic function necessarily remains invariant from one group of examinees to the next, at least among those groups used in defining the complete latent space. This means that *any parameter describing the item characteristic function is an invariant item parameter*.

For binary items, the item characteristic function specifies exactly how the observed responses of a population of examinees depend on the latent traits  $\theta_1, \theta_2, \dots, \theta_k$ . Consequently the item characteristic function is a key concept for making inferences in the reverse direction, that is, for making inferences about unobservable latent traits from the observed item responses. Making such inferences is, as we have said, a basic purpose of mental testing.

Of course the item characteristic functions cannot be observed directly for the simple reason that  $\boldsymbol{\theta}$  is unobservable. If certain assumptions can be made about the shape of these functions, however, then the remaining information of interest about the functions can be inferred from the examinees' responses to the test items. Some models for doing this, some of the practical difficulties involved, and the important implications for test theory and practice if these difficulties are overcome are the topics that constitute the subject matter of the present chapter and the four that follow it. The practical-minded reader is cautioned that immediate use of the results obtained will be computationally difficult at this stage, and also possibly hazardous unless the appropriateness of the model has been investigated.

### 16.3 Local Independence

The item characteristic function is the regression of item score on the latent variables  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ . The practical problem of specifying  $\boldsymbol{\theta}$  is part of the problem of choosing an appropriate model. Such problems are briefly treated in Section 16.11. Until then we shall be concerned with setting up the mathematical model, not with the question of its applicability to any particular set of actual data.

Whenever several items are to be dealt with simultaneously, the assumption of *local independence* is usually considered necessary for effective theoretical

work with item characteristic functions. *Local independence means that within any group of examinees all characterized by the same values  $\theta_1, \theta_2, \dots, \theta_k$ , the (conditional) distributions of the item scores are all independent of each other.* This in no way suggests that item scores are unrelated to each other for the total group of examinees. What it means is that item scores are related to each other only through the latent variables  $\theta_1, \theta_2, \dots, \theta_k$ . We shall show below that local independence is an automatic consequence of the proper choice of  $\theta_1, \theta_2, \dots, \theta_k$ .

The formal definition of local independence for items  $g = 1, 2, \dots, n$  is that for any fixed values of  $\theta_1, \theta_2, \dots, \theta_k$ , the joint distribution  $f$  of item scores  $y_{g*}$  is equal to the product of the marginal distributions  $f_g$ :

$$f(y_{1*}, y_{2*}, \dots, y_{n*} | \boldsymbol{\theta}) = \prod_{g=1}^n f_g(y_{g*} | \boldsymbol{\theta}). \quad (16.3.1)$$

To understand this better, consider one implication of (16.3.1). For  $g = 2, 3, \dots, n$ , (16.3.1) becomes

$$f(y_{2*}, y_{3*}, \dots, y_{n*} | \boldsymbol{\theta}) = \prod_{g=2}^n f_g(y_{g*} | \boldsymbol{\theta}).$$

Divide this into (16.3.1) to obtain

$$h_1(y_{1*} | \boldsymbol{\theta}; y_{2*}, y_{3*}, \dots, y_{n*}) = f_1(y_{1*} | \boldsymbol{\theta}). \quad (16.3.2)$$

This shows that *under local independence, the conditional distribution  $h_1$  of  $y_{1*}$  for fixed values of  $\boldsymbol{\theta}, y_{2*}, y_{3*}, \dots, y_{n*}$  does not depend on  $y_{2*}, y_{3*}, \dots, y_{n*}$ .* This result holds, of course, no matter which item is chosen as item 1.

It may also be shown, conversely, that if (16.3.2) holds for each item, then (16.3.2) implies (16.3.1). Thus either equation constitutes a definition of *local independence*.

Suppose (16.3.2) did not hold. Then for a group of examinees all having the same  $\boldsymbol{\theta}$  and all having the same  $y_{g*}$  for  $g = 2, 3, \dots, n$ , the conditional distribution of  $y_{1*}$  would be different than for other groups of examinees with the same  $\boldsymbol{\theta}$  but different  $\{y_{g*}\}$ . But this cannot occur in the complete latent space, since that space is defined (see preceding section) so that the distribution of  $y_{1*}$  is the same for all pertinent groups of examinees. *The assumption of local independence is thus equivalent to the assumption that the  $\theta_1, \theta_2, \dots, \theta_k$  under consideration span the complete latent space.* Local independence is assumed throughout the present chapter.

For binary items, the assumption of local independence can be written

$$\text{Prob}(U_{1*} = u_1, U_{2*} = u_2, \dots, U_{n*} = u_n | \boldsymbol{\theta}) = \prod_{g=1}^n \text{Prob}(U_{g*} = u_g | \boldsymbol{\theta}), \quad (16.3.3)$$

where  $u_g$  is equal to zero or one. Let us denote the conditional probability for

given  $\boldsymbol{\theta}$  of a correct answer to binary item  $g$  by  $P_g(\boldsymbol{\theta})$ , or simply by  $P_g$  (not to be confused with the  $P_i$  of Chapter 15):

$$P_g \equiv P_g(\boldsymbol{\theta}) \equiv \text{Prob} (U_{g*} = 1 | \boldsymbol{\theta}). \quad (16.3.4)$$

This function  $P_g(\boldsymbol{\theta})$  is the item characteristic function for a binary item; that is, it is the regression of item score on  $\boldsymbol{\theta}$ . The frequency distribution of a binary item score for fixed  $\boldsymbol{\theta}$  can be written

$$f_g(u_g | \boldsymbol{\theta}) \equiv P_g^{u_g} Q_g^{1-u_g}, \quad (16.3.5)$$

where  $Q_g \equiv 1 - P_g$ . Equation (16.3.5) is a compact way of writing

$$f_g(u_g | \boldsymbol{\theta}) \equiv \begin{cases} P_g & \text{if } u_g = 1, \\ Q_g & \text{if } u_g = 0. \end{cases} \quad (16.3.6)$$

The pattern of item response, to be denoted by the vector

$$\mathbf{V}_* \equiv (U_{1*}, U_{2*}, \dots, U_{n*})',$$

is a vector random variable. By (16.3.5) and (16.3.3), the conditional distribution of  $\mathbf{V}_*$  for given  $\boldsymbol{\theta}$  is

$$P(\mathbf{v} | \boldsymbol{\theta}) \equiv \prod_{g=1}^n P_g^{u_g} Q_g^{1-u_g}. \quad (16.3.7)$$

*If, for some population of examinees,  $\boldsymbol{\theta}$  has a frequency distribution denoted by  $g(\boldsymbol{\theta})$ , then  $P(\mathbf{v})$ , the (unconditional) distribution of  $\mathbf{V}_*$  for the total group of examinees, is given by the basic equation*

$$P(\mathbf{v}) \equiv \int g(\boldsymbol{\theta}) \prod_{g=1}^n P_g^{u_g} Q_g^{1-u_g} d\boldsymbol{\theta}, \quad (16.3.8)$$

where the integration is over the  $k$ -dimensional set corresponding to the elements of  $\boldsymbol{\theta}$ . Since we can observe a sample drawn from  $P(\mathbf{v})$ , we can use this equation to make inferences about the unknown distribution  $g$  of the unobservable variables  $\boldsymbol{\theta}$ , provided that the function  $P_g(\boldsymbol{\theta})$  is known.

Lazarsfeld's pioneering work in latent structure analysis (1959, 1960, 1961) is also based on the assumption of local independence. The main difference between his work and that presented here is that Lazarsfeld tends to study all possible response patterns  $\mathbf{v}$ . With dichotomous items, there may be  $2^n$  of these. In mental test theory, where  $n$  is often at least 25, it is usually necessary to try to summarize the information on the examinee's answer sheet by using one or more test scores, such as the number of right answers  $X$ . The reader interested in other developments of latent structure analysis should refer to the brief discussions in Chapters 17 and 24 and to Anderson (1959), Madansky (1960), McDonald (1962, 1967, in press), and Meredith (1965).

#### 16.4 Item-Test Regression

How does one know the form of an item characteristic function  $P_g(\theta)$ ? If a test is composed of sufficiently homogeneous items, for example, of vocabulary, spelling, or some one type of spatial item, we commonly suppose that a sufficiently close representation can be obtained if we assume that  $k = 1$ , that is, if we assume that the complete latent space is one-dimensional. In this case, as will be shown in Section 16.13 and Chapter 18, the usual total test score itself may provide a useful, though less than perfectly reliable, measure of (a monotonic function of)  $\theta_1$ .

Figure 16.4.1(a) and (b) show the empirically determined regressions of item score on test score, two dichotomous verbal items being shown in each plot. Observed percentage of correct answers (= mean item score) is plotted against score on an 89-item vocabulary test. Figure 16.4.1(c) through (f) similarly show observed regression of item score against score on a 59-item mathematics test for each of six mathematics items. Each curve represents the responses of 103,275 examinees. Points that would be based on the performances of less than 50 examinees are not plotted at all.

For the present, we shall mention only two generalizations from these curves:

1. With minor local exceptions, the curves are monotonic nondecreasing.
2. The curves have such a variety of different shapes that it is unlikely that they can be readily fitted by a simple two-parameter family of mathematical functions.

Note that Fig. 16.4.1 shows item-test regressions, *not* item characteristic functions; that is, the base line is the total test score  $x$ , not the latent trait  $\theta_1$ . It has been common to think of an item-test regression as a close approximation to an item characteristic function. However, not only is  $x$  unreliably measured, but the unit of measurement provided by the score scale on test  $X$  is peculiar to the particular test administered. In other words, examinees who differ by one score unit on test  $X$  will tend to differ by more or less than one score unit on a different test of the same trait. Thus a single item will have differently shaped regressions on different tests measuring the same  $\theta$ . This makes it undesirable to use the  $x$  of any particular test as the base line for a general treatment of item characteristic functions.

The following theorem, and especially its corollary, shows that item-test regressions are closely related to the score scale provided by the test. Let  $M_{g|z}$  be the sample (or population) regression of  $Y_{ga}$ , the score on item  $g$ , on total test score:

$$M_{g|z} \equiv \frac{1}{N_z} \sum_{a|z} y_{ga}, \quad z = 0, \frac{1}{n}, \frac{2}{n}, \dots, 1, \quad (16.4.1)$$

where  $z$  is the *total proportion-correct test score*

$$z_a \equiv \frac{1}{n} \sum_{g=1}^n y_{ga}, \quad (16.4.2)$$

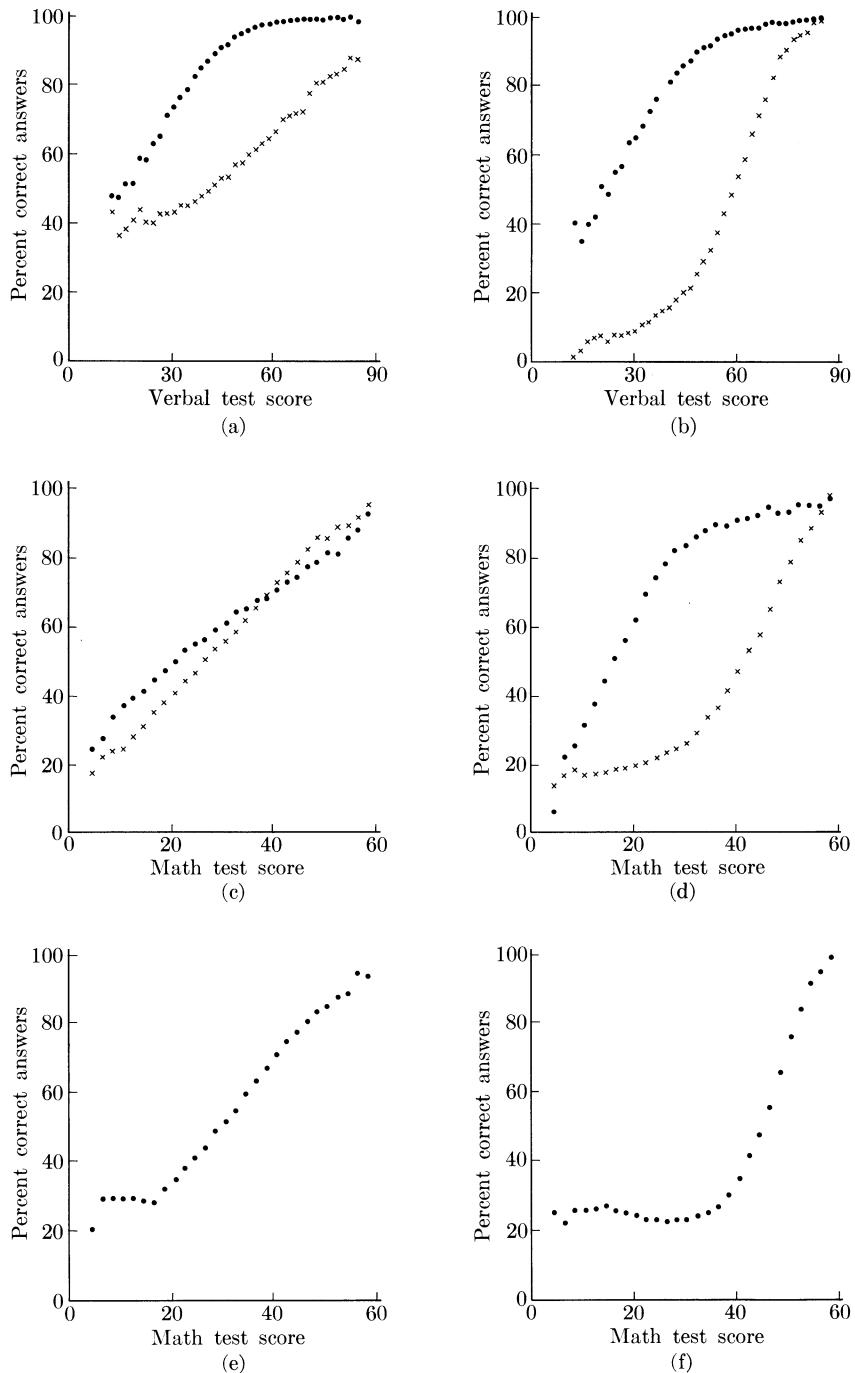


FIG. 16.4.1. Item-test regressions.

that is, the proportion of the  $n$  items answered correctly by examinee  $a$ ; and where  $\sum_{a|z}$  denotes summation over those  $N_z$  examinees in the sample whose test score is  $z_a = z$ .

**Theorem 16.4.1.** *The average, over all  $n$  items, of the sample (or population) item-test regressions falls along a straight line through the origin with  $45^\circ$  slope; that is, for every value of  $z$ ,*

$$\frac{1}{n} \sum_{g=1}^n M_{g|z} \equiv z, \quad z = 0, \frac{n}{1}, \frac{n}{2}, \dots, 1. \quad (16.4.3)$$

*Proof.* Equation (16.4.3) is obtained by averaging (16.4.1) over all examinees whose score is a fixed value  $z_a = z$  and then applying (16.4.2).  $\square$

**Corollary.** If all  $n$  items have identical sample (population) statistics, the sample (population) item-test regression for each item is  $M_{g|z} \equiv z$ ; that is, the regression falls along a straight line through the origin with  $45^\circ$  slope.

It is worth noting that the theorem and corollary hold even if the items are all uncorrelated (or even negatively correlated) with each other. On the average, the item-test regressions have a slope of 1 even though the items do not have any latent trait in common!

As we have previously pointed out, an item-test regression falls short of providing a convenient item characteristic function, not only because of the errors of measurement in  $x$ , but also because  $x$  is determined by the score scale for the particular test in which the item happens to be included and the score scale therefore varies from one test to another. Because of this and because of Theorem 16.4.1, the mathematical form of an item-test regression is complicated rather than simple; it depends not merely on the item itself but also on the characteristics of all the other items in the test.

Levine (1967) has developed a “uniform systems analysis” designed to provide a method for approximating the shape of item characteristic curves for sets of items. His approach assumes only that there exists a scaling for  $\theta_1$  such that the curves for the items form a uniform system, that is, are the same except for a linear transformation of  $\theta_1$ . His theory also provides an approximate analytic determination of this scaling of  $\theta_1$ . Because of its recency and complexity, we shall not describe this approach in detail here. The following sections develop a less general approach to the estimation of item characteristic functions.

## 16.5 The Normal Ogive Model

This chapter is not intended to familiarize the reader completely with all aspects of the normal ogive model. Instead, we shall use this model to illustrate the more general model introduced in the first sections and to provide some background for the logistic models to be treated in greater detail in Chapters 17

through 20. The mathematics of the logistic model are simpler than those of the normal ogive model.

The normal ogive model is intended for use with binary items that have only one latent variable in common. In this case, the item characteristic function is often called the *item characteristic curve*. We denote the latent variable by  $\theta$ . We shall see that the scale used for  $\theta$  is in theory determined, up to a linear transformation, by the assumptions of the model.

The basic assumptions of the normal ogive model are:

1. The latent-variable space is one-dimensional ( $k = 1$ ).
2. The metric for  $\theta$  can be chosen so that the characteristic curve for each item  $g = 1, 2, \dots, n$  (the regression of item score on  $\theta$ ) is the normal ogive

$$P_g(\theta) \equiv P_g(\theta, a_g, b_g) \equiv \Phi[L_g(\theta)] \equiv \int_{-\infty}^{L_g(\theta)} \varphi(t) dt = \int_{-L_g(\theta)}^{\infty} \varphi(t) dt, \quad (16.5.1)$$

where

$$L_g(\theta) \equiv a_g(\theta - b_g) \quad (16.5.2)$$

is a linear function of  $\theta$  involving two item parameters  $a_g$  and  $b_g$ , and  $\varphi(t)$  is the normal frequency function.

Several such characteristic curves are illustrated in Fig. 16.5.1 and 16.5.2. We note that  $P_g(\theta)$  is not used here as a cumulative distribution function, although it has the mathematical properties of a cdf. Other forms that we shall consider for  $P_g(\theta)$  do not have the properties of a cdf.

An appreciation of the qualitative role of the item parameters in such models will be helpful. Figure 16.5.1 illustrates the item characteristic curves for several hypothetical items. Item 1, with  $a_1 = 0$  and  $b_1 = -0.25$ , has item characteristic curve  $P_1(\theta) = \Phi(0) = 0.5$ , a constant independent of  $\theta$ . A response to any item with  $a_g = 0$  is worthless as information about the value of an examinee's ability  $\theta$ , since each examinee has the same probability of giving the response  $u_g = 1$ , regardless of his ability.

If  $a_g > 0$ , then  $\Phi[a_g(\theta - b_g)]$  is strictly increasing in  $\theta$ . The corresponding  $u_g$  is an indicant and a measure of  $\theta$ .

Item 2, with  $a_2 = 0.01$  and  $b_2 = 18$ , has approximately the same property as item 1 over the range of  $\theta$  illustrated, since  $a_2$  is so near zero that  $P_2(\theta) \doteq 0.43$  for  $-3 < \theta < 3$ . Item 3, with  $a_3 = 100$  and  $b_3 = 1$ , has  $P_3(\theta) = \Phi(100\theta - 100)$ . This determines that  $P_3(\theta) < 0.01$  for all  $\theta < 0.9$ , and within this range of ability  $P$  is approximately constant again, as it is also when it is within the range  $\theta > 1.1$ , where  $P_3(\theta) > 0.99$ . This item alone would provide almost sure discrimination between examinees below 0.9 and examinees above 1.1 because, if we interpret  $y_3 = 0$  as indicating  $\theta < 0.9$  and  $y_3 = 1$  as indicating  $\theta > 1.1$ , then the probability of an erroneous indication is less than 0.01 for each possible ability  $\theta$ . The uncommonly high value  $a_3 = 100$  de-

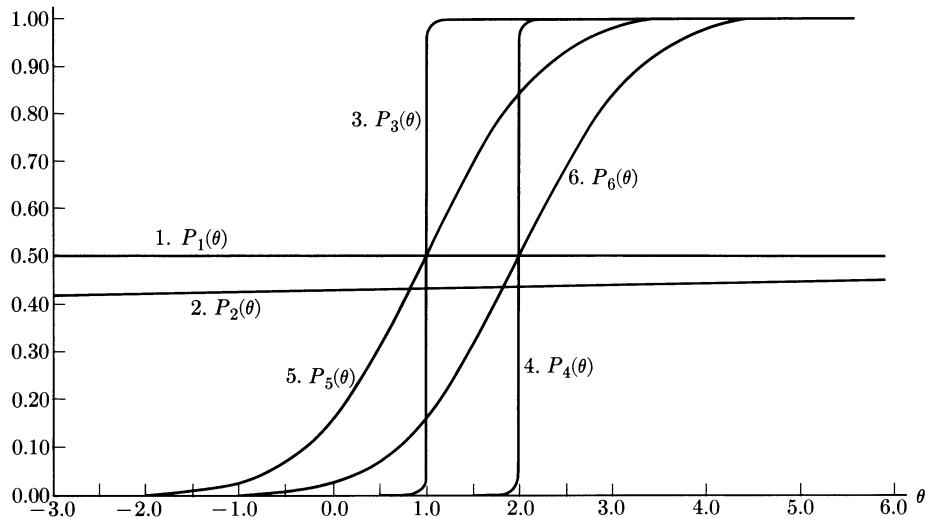


FIG. 16.5.1. Several hypothetical item characteristic curves of the normal ogive (and logistic) forms.

termines that  $P_3$  rises steeply in some small interval and also, consequently, that it will be very near zero to the left of that interval and very near one to the right of that interval. These considerations illustrate, by extreme cases, that  $a_g$  is a parameter that indicates the quality or value of an item in the basic sense of the amount of information the item provides about  $\theta$ . This notion of information requires further refinement: As we have seen with item 3, in general it requires qualification concerning the ranges of  $\theta$ -values where the item does or does not discriminate effectively. We shall call the parameter  $a_g$  the *discriminating power* of the item.

If an item is to be scored so that a “correct” answer gives the examinee a higher score than an “incorrect” answer, then the item will be useful only if the probability of a “correct” answer increases as  $\theta$  increases. Thus we restrict consideration to items that are scored so that  $0 \leq a_g \leq \infty$ . It will ordinarily be assumed that  $a_g$  is finite.

Item 4, with  $a_4 = 100$  and  $b_4 = 2$ , has properties like those of item 3, except that the region of sharp discrimination is shifted from  $\theta = 1$  to  $\theta = 2$ . A test consisting just of items 3 and 4 would give very effective discrimination between three levels of ability: The response pattern  $\mathbf{v}' = (u_3, u_4) = (1, 1)$ , indicating the highest level,  $\theta \geq 2$ , would occur only with probability  $P_3(\theta)P_4(\theta) < 0.01$  for  $\theta < 1.9$ . Similarly,  $\mathbf{v}' = (0, 0)$ , indicating the lowest level,  $\theta \leq 1$ , would occur only with probability  $Q_3(\theta)Q_4(\theta) < 0.01$  for  $\theta > 1.1$ . And  $\mathbf{v}' = (1, 0)$ , indicating the middle level,  $1 < \theta < 2$ , would occur only with probability  $P_3(\theta)Q_4(\theta) < 0.01$  for  $\theta < 0.99$  or  $\theta > 2.1$ . The remaining response pattern  $\mathbf{v}' = (0, 1)$  can also be considered to indicate the middle level; it would occur with negligible probabilities for all values of  $\theta$ .

Items 3 and 4 differ just in their parameters  $b_g$ , and  $P_4(\theta)$  is identical to  $P_3(\theta)$  except for a translation to the right. With items of the normal ogive form, note that for any fixed  $d > 0$ , the maximum of the difference  $P_g(\theta + d) - P_g(\theta - d)$  is attained at  $\theta = b_g$ . This illustrates the fact that  $b_g$  is related to the level of ability at which an item discriminates most effectively. In this sense,  $b_g$  represents the *difficulty level of item g*. Ordinarily it is assumed that  $-\infty < b_g < \infty$ . For the normal ogive model,

$$P_g(b_g) \equiv 0.5 \quad (16.5.3)$$

for any  $b_g$ . Items 3 and 4 have respective difficulty levels  $b_3 = 1$  and  $b_4 = 2$ . Items 5 and 6 have the same respective difficulty levels, namely,  $b_5 = 1$  and  $b_6 = 2$ , and they have a common value  $a_5 = a_6 = 1$ .

On the range illustrated, item characteristic curves 5 and 6 rise moderately rapidly with increase of  $\theta$ , in contrast with curves 1 and 2. However, they rise much less rapidly than do curves 3 and 4 in the neighborhoods of their respective points of inflexion. Each curve of the form  $P_g(\theta) = \Phi[a_g(\theta - b_g)]$  increases approximately linearly in a small interval centered at its point of inflection, and the rate of increase in this interval is approximately proportional to  $a_g$ . Thus  $a_g$  is an index of item discriminating power. More precisely, at  $\theta = b_g$  the derivative of  $P_g(\theta)$  is  $a_g\varphi(0) \equiv a_g/\sqrt{2\pi}$ , which is proportional to  $a_g$ .

To summarize: *The normal ogive item characteristic curve has a point of inflection at  $\theta = b_g$ ; at this point the probability of a correct answer is 0.5, and the slope of the curve is  $a_g/\sqrt{2\pi}$ .*

We can see that items with very high  $a_g$  values, such as items 3 and 4, would be extremely effective for certain discriminations, specifically, for discriminating abilities at least a little below  $b_g$  from values at least a little above  $b_g$ . Notwithstanding this, however, they would be virtually useless for other discriminations, in particular those between any two abilities both a little above  $b_g$  or both a little below  $b_g$ . This is the attenuation paradox already mentioned at the end of Section 15.7. Although items with moderate  $a_g$  values, such as items 5 and 6, are far less effective for discriminations of the first type, we shall find that they do have some discriminating power over an interval of greater or lesser width. Eventually we shall have occasion to examine this matter in detail. Here it suffices to point out that with a test consisting of a moderately large number of items with the same parameters as item 5, the simple test score  $z$  would be approximately equal to  $P_5(\theta)$  with high probability for each  $\theta$ ; hence such a test score would discriminate rather effectively between all appreciably different values of  $\theta$  over a wide range. Such a possibility is most important for applications because the items usually available in practice are found to have only moderate  $a_g$  values. In contrast, a test containing the same number of items with parameters identical with those of items 3 or 4 would fail to give equally effective discrimination between such values as  $\theta = 1.4$  and  $\theta = 1.6$ . The fact is that increasing the  $a_g$  values may decrease the overall

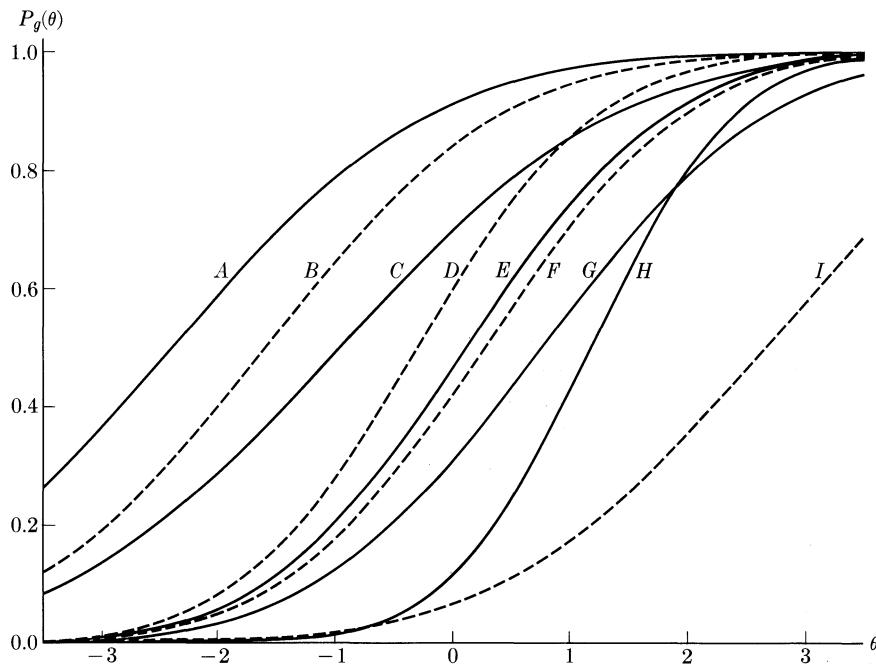


FIG. 16.5.2. Item characteristic curves.

discriminating power of a test and the overall correlation of test score with  $\theta$ . If this appears paradoxical, it is because the role of the parameters  $a_g$  is considered too simply, as will be pointed out in Chapter 20.

Whenever any single item characteristic curve is a monotonic increasing function of  $\theta$ , it is always possible and permissible to transform  $\theta$  monotonically so that the characteristic curve becomes a normal ogive. In general, different transformations would ordinarily be needed to achieve this for different items. Thus the assumption that the item characteristic curves of  $n$  items can all be represented simultaneously by normal ogives is a restrictive assumption, the utility of which must be investigated for any actual set of items. If the restrictive assumption holds to a satisfactory approximation, then it determines the metric for  $\theta$  up to an arbitrary linear transformation.

Figure 16.5.2 shows normal ogives estimated from actual test data for nine free-response items. The nine items were chosen to cover a wide range of difficulty, so there is less crowding and crossing of curves than would be found in a typical test. We have followed the common practice of choosing the origin and unit of measurement for  $\theta$  so that the estimated mean of  $\theta$  for the sample of examinees is zero and the estimated variance is one.

Lawley (1943, 1944) was the first to develop a normal ogive model for  $n$  items,  $n > 2$ . This model has been used in theoretical work by many writers,

including Brogden (1946), Tucker (1946), Cronbach and Warrington (1952), Lord (1952, 1953a, 1953b), Cronbach and Merwin (1955), Cronbach and Azuma (1962), and Paterson (1962). It has been used in practical test construction by Indow and Samejima (1962, 1966). Bock (1967) has reported new methods for fitting the model to test data.

### 16.6 Conditions Leading to the Normal Ogive Model

Equation (16.5.1) may be taken simply as a basic assumption, the utility of which can be investigated for a given set of data (albeit with considerable difficulty). Alternatively (16.5.1) can be inferred from other, possibly more plausible assumptions. We shall outline one way of doing this, a way that some theorists find interesting and others do not.

Suppose that for each item  $g = 1, 2, \dots, n$ , there is a continuous random variable  $Y'_{ga}$  or  $Y'_g$  that varies over the population  $\mathcal{P}$  of examinees and that has the following properties:

1. Whenever  $Y'_{ga}$  is greater than some constant  $\gamma_g$  characterizing the item, examinee  $a$  answers item  $g$  correctly. We denote this by  $U_{ga} = 1$ .
2. Whenever  $Y'_{ga} < \gamma_g$ , he answers it incorrectly. We denote this by  $U_{ga} = 0$ .

Thus  $Y'_g$  may be thought of as the hypothetical trait of the examinee that determines whether he will answer item  $g$  correctly.

3. The regression of  $Y'_g$  on  $\theta$  is linear.
4. The conditional distribution of  $Y'_g$  given  $\theta$  is normal, with mean denoted by  $\mu'_{g|\theta}$  and variance denoted by  $\sigma'^2_{g|\theta}$ .
5. The variance  $\sigma'^2_{g|\theta}$  is independent of  $\theta$ .

The entire situation is illustrated by Fig. 16.6.1.

The conditional frequency distribution of  $Y'_g$  for given  $\theta$  is therefore

$$\frac{1}{\sigma'_{g|\theta} \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma'^2_{g|\theta}} (y'_g - \mu'_{g|\theta})^2 \right]. \quad (16.6.1)$$

It follows from (16.6.1) that

$$\text{Prob}(Y'_g > \gamma_g | \theta) = \int_{\gamma_g}^{\infty} \frac{1}{\sigma'_{g|\theta} \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma'^2_{g|\theta}} (y'_g - \mu'_{g|\theta})^2 \right] dy'_g. \quad (16.6.2)$$

Make the transformation  $T \equiv (Y'_g - \mu'_{g|\theta})/\sigma'_{g|\theta}$  and obtain

$$\begin{aligned} P_g(\theta) &\equiv \mathcal{E}(U_g | \theta) \equiv \text{Prob}(U_g = 1 | \theta) = \text{Prob}(Y'_g > \gamma_g | \theta) \\ &= \text{Prob}\left(T > \frac{\gamma_g - \mu'_{g|\theta}}{\sigma'_{g|\theta}}\right) = \Phi\left(\frac{\mu'_{g|\theta} - \gamma_g}{\sigma'_{g|\theta}}\right). \end{aligned} \quad (16.6.3)$$

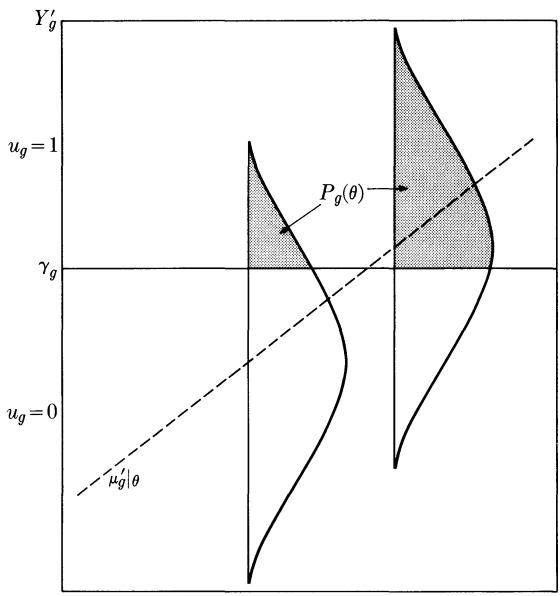


FIG. 16.6.1. Hypothetical relation between item response  $u_g$  and latent trait  $\theta$ .

Since  $\mu'_{g|\theta}$  is a linear function of  $\theta$ , and since  $\sigma'_{g,\theta}$  is constant, this result is the same as (16.5.1). Therefore we have shown that if the stated conditions hold for each of  $n$  items, then the characteristic function is a normal ogive for every item.

Note that this derivation involves no assumption about the frequency distribution of  $\theta$ , or of  $Y'_g$ , over the total group of examinees. As we pointed out in Section 16.2, the assumption that  $k = 1$ , that is, that the complete latent space is one-dimensional, is equivalent to the assumption that the item characteristic function is the same for any total group of examinees that is of interest, regardless of the frequency distribution of  $\theta$  in this total group.

The discussion of the normal ogive characteristic curve in Section 16.5 and in the present section has made assumptions about unobservable quantities. Is it possible to state the assumptions in terms of observable quantities? We shall take a step in this direction when we state the main theorem in Section 16.8. We shall give a preliminary definition and lemma in Section 16.7.

### 16.7 Correlation Matrix with One Common Factor\*

Let  $\|\rho'_{gh}\|$  be the intercorrelation matrix among any  $n$  variables. Suppose that  $\rho'^{2}_g$ ,  $g = 1, 2, \dots, n$ , can be found with  $0 < \rho'^{2}_g \leq 1$ , such that when the  $\rho'^{2}_g$  are substituted for the main-diagonal entries of  $\|\rho'_{gh}\|$ , the resulting matrix is

\* While the abbreviated treatment of factor analysis given in this section is sufficient for subsequent developments in this chapter, some readers may wish to study Sections 24.1 through 24.3 at this time.

of rank 1. In such a case, the variables are said to have exactly *one common factor* and the  $\rho_g'^2$  are called *communalities*. Otherwise, the variables are said to have more than one common factor. The nature of the common factor will be clarified in the following development. We shall give a more complete introduction to factor analysis in Chapter 24.

When  $n = 2$ , we see immediately from the definition given above that the variables necessarily have just one common factor whenever  $\rho_{12} \neq 0$ . In the remainder of this section, we consider only the case where  $n > 2$ .

The rank of a square matrix is defined to be the number of linearly independent rows (or columns) of the matrix. Thus, in a matrix that has just one common factor, the elements in any one row (column) are proportional to the corresponding elements of any other specified row (column) except where diagonal elements are involved. In other words,

$$\rho'_{gh} = C_{gi} \rho'_{ih}, \quad g, h, i = 1, 2, \dots, n, \quad h \neq g, i, \quad (16.7.1)$$

where  $C_{gi}$  is constant for any given  $g$  and  $i$ . For example, the following matrix has just one common factor:

$$\begin{vmatrix} 1.00 & 0.40 & 0.10 & 0.20 \\ 0.40 & 1.00 & 0.20 & 0.40 \\ 0.10 & 0.20 & 1.00 & 0.10 \\ 0.20 & 0.40 & 0.10 & 1.00 \end{vmatrix}.$$

We may verify this by substituting 0.20, 0.80, 0.05, and 0.20 for the diagonal elements, in that order. The trivial case where all off-diagonal correlations in a row (column) are zero is ruled out of consideration in the following discussion.

We may obtain an equation equivalent to (16.7.1) by substituting  $j$  for  $h$ :

$$\rho'_{gj} = C_{gi} \rho'_{ij}, \quad j \neq g, i. \quad (16.7.2)$$

Divide (16.7.1) by (16.7.2) and clear of fractions to obtain Spearman's famous *tetrad difference equation*

$$\rho'_{gh} \rho'_{ij} = \rho'_{gj} \rho'_{ih}, \quad g, h, i, j \text{ unequal}, \quad (16.7.3)$$

which holds for any set of four distinct variables that have just one common factor.

Multiply (16.7.3) by  $\rho'_{gi}/\rho'_{ij} \rho'_{ih}$  and obtain an important relationship between two *triads*:

$$\rho'_{gi} \rho'_{gh}/\rho'_{ih} = \rho'_{gi} \rho'_{gj}/\rho'_{ij}, \quad g, h, i, j \text{ unequal}. \quad (16.7.4)$$

This line of proof shows that for fixed  $g$ , all triads of the form shown in (16.7.4) are equal whenever there is just one common factor. Denote their common value by  $\rho_g'^2$ , so that

$$\rho_g'^2 \equiv \rho'_{gi} \rho'_{gj}/\rho'_{ij}, \quad g, i, j \text{ unequal}. \quad (16.7.5)$$

Exchange of  $g$  and  $i$  in (16.7.5) produces a formula for  $\rho_i'^2$ . From this and (16.7.5) we see that

$$\rho_g' \rho_i' = \rho_{gi}', \quad g, i, j \text{ unequal.} \quad (16.7.6)$$

Thus the entire matrix  $\|\rho_{gi}'\|$  can be computed from the vector  $\{\rho_g'\}$ . We see that if the  $\rho_g'^2$  are substituted for the diagonal elements  $\rho_{gg}' = 1$  of  $\|\rho_{gi}'\|$ , then the resulting matrix is of unit rank. Thus the communalities  $\rho_g'^2$  defined at the beginning of the present section are the same as the  $\rho_g'^2$  obtained from (16.7.5). Each of these results is part of the meaning of the statement that the  $y'$ -variables have only one common factor.

It is important to note that the common factor of the random variables  $y_g', g = 1, 2, \dots, n$ , is itself not a uniquely defined variable. In fact any variable  $\theta$  whose correlation with each  $y_g'$  is  $\rho_g'$ ,  $g = 1, 2, \dots, n$ , can be called "the common factor of the  $y_g'$ ". The reader can verify that if  $\theta$  and  $E_g, g = 1, 2, \dots, n$ , are uncorrelated random variables, each with unit variance, then the correlation matrix between the  $Y_g$  defined by

$$Y_g \equiv \rho_g \theta + E_g \sqrt{1 - \rho_g^2}, \quad g = 1, 2, \dots, n, \quad (16.7.7)$$

is of rank 1 when the  $\rho_g^2$  are placed in the diagonal. Thus  $\theta$  is a common factor of the  $Y$ . Equation (16.7.7) thus represents each observable variable  $Y_g$  as a weighted sum of a (latent) common variable  $\theta$  and an uncorrelated "unique" variable  $E_g$ . This type of representation is basic to all factor analysis.

It can be shown from (16.7.6) that if  $\|\rho_{gh}'\|$  has just one common factor, then  $\rho$ , the determinant of  $\|\rho_{gh}'\|$ , is

$$\rho \equiv G \prod_{g=1}^n K_g^2, \quad (16.7.8)$$

where

$$K_g^2 \equiv 1 - \rho_g'^2 \quad (16.7.9)$$

and

$$G \equiv 1 + \sum_{g=1}^n \frac{\rho_g'^2}{K_g^2} = 1 - n + \sum_{g=1}^n \frac{1}{K_g^2}.$$

The following lemma may now be readily verified by multiplying  $\|\rho_{gh}'\|$  by its inverse.

**Lemma 16.7.1.** If  $\|\rho_{gh}'\|$  has just one common factor, then its inverse  $\|\rho'^{gh}\|$  has off-diagonal elements

$$\rho'^{gh} = -\rho_g' \rho_h' / GK_g^2 K_h^2$$

and diagonal elements

$$\rho'^{gg} = 1/K_g^2 - \rho_g'^2/GK_g^4.$$

### 16.8 A Sufficient Condition for Normal Ogive Item Characteristic Curves\*

One purpose of the following theorem is to show that there is nothing implausible about the assumption that the items in a test all have normal ogive characteristic curves.

**Theorem 16.8.1.** If

a) the frequencies constituting the multivariate distribution of  $U_{1*}, U_{2*}, \dots, U_{n*}$  for some specified group of examinees could have arisen from some multivariate normal distribution of  $Y'_{1*}, Y'_{2*}, \dots, Y'_{n*}$  by dichotomizing each  $Y'$ -variable, and

b) the  $Y'$ -variables have just one common factor,

then for this group of examinees, the multivariate distribution of  $U_{1*}, U_{2*}, \dots, U_{n*}$  is consistent with the assumptions that

- i) the complete latent space is one-dimensional,
- ii) all item characteristic curves are normal ogives,
- iii) local independence holds for the  $n$  items, and
- iv) the latent trait  $\theta$  is normally distributed.

*Proof.* Let  $\mathbf{v} = (u_1, u_2, \dots, u_n)$  denote any pattern of  $n$  binary responses ( $u_g = 0$  or  $1$ ), and let  $p(\mathbf{v})$  denote the unconditional probability of  $\mathbf{v}$  for the given population of examinees. The first condition of the theorem states that

$$p(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^n \rho}} \int_{\gamma'} \cdots \int \exp \left[ -\frac{1}{2} \left( \sum_g \sum_h \rho'^{gh} y'_g y'_h \right) \right] \prod_{g=1}^n dy'_g, \quad (16.8.1)$$

where  $\gamma'$  is an appropriate  $n$ -dimensional rectangular region. It will be notationally convenient to work first with  $u_1 = u_2 = \cdots = u_n = 1$ . Here

$$p(1, 1, \dots, 1) = \frac{1}{\sqrt{(2\pi)^n \rho}} \int_{\gamma_1}^{\infty} \int_{\gamma_2}^{\infty} \cdots \int_{\gamma_n}^{\infty} \exp \left[ -\frac{1}{2} \left( \sum_g \sum_h \rho'^{gh} y'_g y'_h \right) \right] \prod_{g=1}^n dy'_g. \quad (16.8.2)$$

Multiply (16.8.2) by the identity

$$1 \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx,$$

to obtain

$$\begin{aligned} p(1, 1, \dots, 1) &= \frac{1}{\sqrt{(2\pi)^{n+1} \rho}} \int_{-\infty}^{\infty} \int_{\gamma_1}^{\infty} \int_{\gamma_2}^{\infty} \\ &\quad \cdots \int_{\gamma_n}^{\infty} \exp \left[ -\frac{1}{2} \left( x^2 + \sum_g \sum_h \rho'^{gh} y'_g y'_h \right) \right] \prod_{g=1}^n dy'_g dx. \end{aligned} \quad (16.8.3)$$

---

\* Reading of this section may be omitted without loss of continuity.

Define  $\theta$  by the equation

$$\chi \equiv \sqrt{G} \theta - \frac{1}{\sqrt{G}} \sum_g \frac{\rho'_g y'_g}{K_g^2}.$$

Square, rearrange, and apply Lemma 16.7.1 to obtain

$$\chi^2 + \sum_g \sum_h \rho'^{gh} y'_g y'_h \equiv \theta^2 + \sum_g \frac{(y'_g - \rho'_g \theta)^2}{K_g^2}.$$

Using this and (16.7.8), we write Eq. (16.8.3) as

$$p(1, 1, \dots, 1) = \int_{-\infty}^{\infty} \varphi(\theta) \prod_{g=1}^n \left\{ \int_{\gamma_g}^{\infty} \frac{1}{\sqrt{2\pi} K_g} \exp \left[ -\frac{1}{2K_g^2} (y'_g - \rho'_g \theta)^2 \right] dy'_g \right\} d\theta.$$

After the change of variables

$$t_g \equiv \frac{y'_g - \rho'_g \theta}{K_g}, \quad g = 1, 2, \dots, n,$$

we find that

$$p(1, 1, \dots, 1) = \int_{-\infty}^{\infty} \varphi(\theta) \prod_{g=1}^n \left[ \int_{(\gamma_g - \rho'_g \theta)/K_g}^{\infty} \varphi(t_g) dt_g \right] d\theta. \quad (16.8.4)$$

If

$$a_g \equiv \rho'_g / \sqrt{1 - \rho'^2_g} \quad \text{and} \quad b_g \equiv \gamma_g / \rho'_g, \quad (16.8.5)$$

then the term in brackets in (16.8.4) is the same as  $P_g(\theta)$  in (16.5.1), so that finally

$$p(1, 1, \dots, 1) = \int_{-\infty}^{\infty} \varphi(\theta) \prod_{g=1}^n P_g(\theta) d\theta.$$

The same derivation for any response pattern  $\mathbf{v}$  yields the general result

$$p(\mathbf{v}) = \int_{-\infty}^{\infty} \varphi(\theta) \prod_{g=1}^n [P_g(\theta)]^{u_g} [Q_g(\theta)]^{1-u_g} d\theta. \quad (16.8.6)$$

This result is the same as (16.3.8) for normally distributed  $\theta$ .  $\square$

We have already pointed out at the end of Section 16.6 that the item characteristic function of an item remains the same from one group of examinees to another. Thus if  $n$  items have normal ogive characteristic curves when the group of examinees has a normal distribution of  $\theta$ , as in (16.8.6), the items will still have the same normal ogive characteristic curves for any other group of examinees, even though  $\theta$  is not normally distributed. The conditions stated in Theorem 16.8.1 require a normal distribution of  $\theta$ . Thus these sufficient conditions are very far from being necessary conditions.

When  $n = 2$ , any set of item-response data will always satisfy condition (a) of the theorem. This is implicit in the fact that every two  $\times$  two table always uniquely determines a tetrachoric correlation coefficient  $\rho'_{gh}$  (Section 15.8). If  $\rho'_{gh} \neq 0$ , then condition (b) of the theorem is satisfied (Section 16.7). When  $n > 2$ , neither condition (a) nor (b) need be satisfied. For example, the accompanying two  $\times$  two  $\times$  two table violates condition (a). Symmetry considerations tell us that this set of data could not arise by the triple dichotomization of a trivariate normal distribution.

### 16.9 Normal Ogive Parameters: Item Difficulty

Clearly  $a_g$  and  $b_g$  in (16.5.2) are parameters for the normal ogive of binary item  $g$ . If all assumptions are met, they should remain invariant from one group of examinees to another. If values of an invariant item parameter are determined by pretesting, these can be used to predict the properties of a test for various groups of examinees.

Practical workers will wish to know how  $a_g$  and  $b_g$  are related to the familiar and commonly used item parameters discussed in Chapter 15. It is the purpose of this section and the following one to derive and discuss these relationships.

Since the parameters of Chapter 15 depend on the characteristics of the group of examinees, whereas  $a_g$  and  $b_g$  do not, it is clear that the relation between the two sets of parameters will be different for different groups. Here we shall spell the relation out only for the convenient case where  $\theta$  is normally distributed. A parallel derivation will be valid for any other specified distribution of  $\theta$ .

Without loss of generality, we may choose the origin and measurement scale for  $\theta$  so that  $\mu_\theta = 0$ ,  $\sigma_\theta = 1$ . First of all, if the item characteristic curve is

$$P_g(\theta) = \int_{a_g(b_g - \theta)}^{\infty} \varphi(t) dt, \quad (16.9.1)$$

what is the proportion of correct answers in a population where  $\theta$  is distributed  $N(0, 1)$ , that is, normally, with zero mean and unit variance?

In (16.9.1), make the transformation

$$y' \equiv t/a_g + \theta,$$

so that

$$P_g(\theta) = a_g \int_{b_g}^{\infty} \varphi[a_g(y' - \theta)] dy'. \quad (16.9.2)$$

The proportion of correct answers to item  $g$  in a group of examinees, that is,

the *difficulty* of item  $g$  for that group, is

$$\begin{aligned}\pi_g \equiv \text{Prob } (u_g = 1) &= \int_{-\infty}^{\infty} P_g(\theta) \varphi(\theta) d\theta \\ &= \frac{a_g}{2\pi} \int_{-\infty}^{\infty} \int_{b_g}^{\infty} \exp \left\{ -\frac{1}{2}[\theta^2 + a_g^2(y' - \theta)^2] \right\} dy' d\theta \\ &= \frac{a_g}{2\pi} \int_{b_g}^{\infty} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2}[(1 + a_g^2)\theta^2 - 2a_g^2y'\theta + a_g^2y'^2] \right\} d\theta dy'.\end{aligned}$$

Complete the square of the exponent and let

$$w \equiv \theta \sqrt{1 + a_g^2} - a_g^2 y' / \sqrt{1 + a_g^2},$$

obtaining

$$\begin{aligned}\pi_g &= \frac{a_g}{2\pi} \int_{b_g}^{\infty} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} \left[ \left( \theta \sqrt{1 + a_g^2} - \frac{a_g^2 y'}{\sqrt{1 + a_g^2}} \right)^2 \right. \right. \\ &\quad \left. \left. + \left( a_g^2 - \frac{a_g^4}{1 + a_g^2} \right) y'^2 \right] \right\} d\theta dy' \\ &= \frac{a_g}{\sqrt{2\pi} \sqrt{1 + a_g^2}} \int_{b_g}^{\infty} \exp \left( -\frac{1}{2} \frac{a_g^2 y'^2}{1 + a_g^2} \right) dy' \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw.\end{aligned}$$

The last integral above is equal to 1.

Finally let  $t \equiv a_g y' / \sqrt{1 + a_g^2}$  to obtain

$$\begin{aligned}\pi_g &= \frac{1}{\sqrt{2\pi}} \int_{\gamma_g}^{\infty} e^{-t^2/2} dt \\ &= \Phi(-\gamma_g),\end{aligned}\tag{16.9.3}$$

where  $\gamma_g$  is defined by

$$\gamma_g \equiv a_g b_g / \sqrt{1 + a_g^2}.\tag{16.9.4}$$

(This last equation is consistent with (16.8.5), which we derived from other assumptions.)

Where  $\theta$  is normally distributed with zero mean and unit variance, (16.9.3) and (16.9.4) show the relationship of  $\pi_g$ , the item difficulty, to the item parameters  $a_g$  and  $b_g$  of (16.9.1). We shall discuss the practical use of  $a_g$  and  $b_g$  in Section 16.11.

## 16.10 Normal Ogive Parameters: Item Discriminating Power

**Correlation of item with latent trait.** Next, given that the item characteristic curve is a normal ogive (16.9.1) and that  $\theta$  is  $N(0, 1)$ , what is the correlation  $\rho_{\theta g}$  between  $\theta$  and  $u_g$ ?

The bivariate distribution of  $\theta$  and  $u_g$  is

$$f(\theta, u_g) \equiv \varphi(\theta) P_g^{u_g} Q_g^{1-u_g}. \quad (16.10.1)$$

The conditional distribution of  $\theta$  for  $u_g = 1$  is therefore

$$h(\theta | u_g = 1) \equiv \frac{1}{\pi_g} \varphi(\theta) P_g(\theta). \quad (16.10.2)$$

The conditional mean of  $\theta$  when  $u_g = 1$  is

$$\mu^+ \equiv \frac{1}{\pi_g} \int_{-\infty}^{\infty} \theta \varphi(\theta) P_g(\theta) d\theta.$$

When we carry this integration out, we find that

$$\mu^+ = \frac{\varphi(\gamma_g)}{\pi_g} \rho'_g, \quad (16.10.3)$$

where  $\gamma_g$  is given by (16.9.4) and where  $\rho'_g$  is defined by

$$\rho'_g \equiv a_g / \sqrt{1 + a_g^2}. \quad (16.10.4)$$

We now find the product moment correlation between  $\theta$  and  $u_g$  by substituting (16.10.3) into the last part of formula (15.5.4) for a point biserial correlation. Since  $\mu_\theta = 0$  and  $\sigma_\theta = 1$ , the result is

$$\rho_{\theta g} = \rho'_g \frac{\varphi(\gamma_g)}{\sqrt{\pi_g(1 - \pi_g)}}. \quad (16.10.5)$$

Now  $\rho_{\theta g}$  is a point biserial correlation coefficient (Section 15.5). Since the right-hand side of (16.10.5) is the same as that of (15.7.1), which gives the relation of a biserial to a point biserial correlation, we see that the parameter  $\rho'_g$  defined by (16.10.4) is the same as the *biserial* correlation  $\rho'_{\theta g}$  between  $\theta$  and  $u_g$ :

$$\rho'_{\theta g} = \rho'_g. \quad (16.10.6)$$

This correlation is a measure of the discriminating power of the item.

By (16.10.4), if  $\theta$  is  $N(0, 1)$ , then

$$a_g = \rho'_g / \sqrt{1 - \rho'^2_g}. \quad (16.10.7)$$

We have obtained this same formula [the left-hand definition of (16.8.5)] from other assumptions. In this context, we see that *when  $\theta$  is normally distributed, the item parameter  $a_g$  is a known monotonic increasing function of the biserial correlation between the item and the latent trait  $\theta$ .*

**Correlation between two items.** Finally, if (16.9.1) holds for items  $g$  and  $h$ , and  $\theta$  is  $N(0, 1)$ , what is the correlation between  $u_g$  and  $u_h$ ?

Starting with the trivariate distribution for  $\theta$ ,  $u_g$ , and  $u_h$ , and integrating out  $\theta$ , we find that if  $\theta$  is normally distributed, then  $\rho'_{gh}$ , the tetrachoric correlation between  $u_g$  and  $u_h$ , is equal to the product of the biserial correlations of item  $g$  and of item  $h$  with the latent trait  $\theta$ :

$$\rho'_{gh} = \rho'_g \rho'_h. \quad (16.10.8)$$

It follows from this, as it did in (16.7.6), that the matrix  $\|\rho'_{gh}\|$  has just one common factor, as given in (16.7.5):

$$\rho'^2_g = \rho'_{gi} \rho'_{gj} / \rho'_{ij}. \quad (16.10.9)$$

If  $\theta$  is  $N(0, 1)$  in the population of examinees tested, we can use (16.10.9) to obtain the value of each item parameter  $\rho'_g$  from the matrix of tetrachoric item intercorrelations, and hence obtain the value of each parameter  $a_g$ ,  $g = 1, 2, \dots, n$ .

We have seen from (16.9.3), (16.9.4), (16.10.4), and (16.10.9) that when  $\theta$  is  $N(0, 1)$ , the parameters  $a_g$  and  $b_g$  are invariant item parameters that have a clear and direct relationship to item difficulty and item intercorrelation.

## 16.11 Practical Use of Normal Ogive Item Parameters

To provide some notion of the nature of the item parameters under discussion here, we list the item parameters for the nine-item test of Fig. 16.5.2 in Table 16.11.1. We might note that it is rare to find values of  $a_g$  as large as 2 in aptitude and achievement testing.

The parameters  $\pi_g$  and  $\rho'_g$  are helpful in thinking about the item characteristic curve and its properties. These parameters are invariant under a linear transformation of the metric (arbitrarily) chosen for  $\theta$ ; in general they are *not* invariant from group to group, however. The parameters  $a_g$  and  $b_g$  are invariant from group to group but they are *not* invariant when the unit and origin arbitrarily selected for  $\theta$  are changed.

**Table 16.11.1**  
Item parameters for ogives shown in Fig. 16.5.2

Curve	$\pi_g$	$\gamma_g$	$\rho'_g$	$a_g$	$b_g$
A	0.096	1.305	0.490	0.562	2.664
B	0.199	0.845	0.717	1.029	1.178
C	0.338	0.418	0.549	0.657	0.761
D	0.434	0.166	0.593	0.736	0.280
E	0.471	0.073	0.595	0.740	0.123
F	0.574	-0.187	0.640	0.833	-0.292
G	0.676	-0.457	0.476	0.541	-0.961
H	0.801	-0.845	0.530	0.625	-1.594
I	0.882	-1.185	0.495	0.570	-2.393

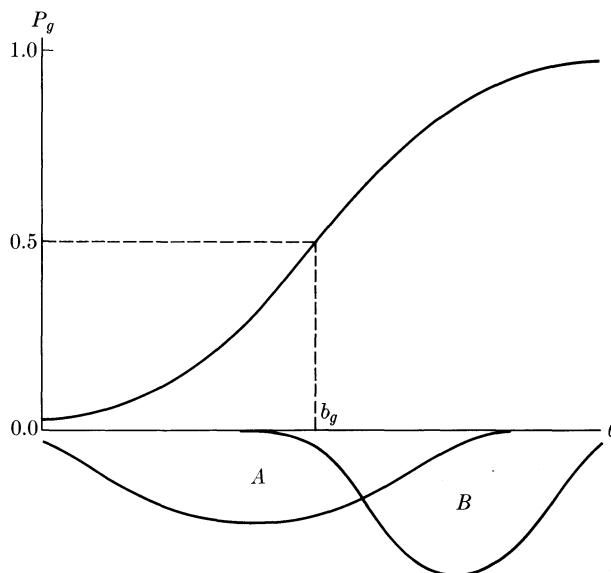


FIG. 16.11.1. Item characteristic curve in relation to two groups of examinees.

These facts are illustrated by Fig. 16.11.1, which shows an item characteristic curve and (upside down) the frequency distributions of  $\theta$  for two groups,  $A$  and  $B$ . The value of  $b_g$  on whatever scale is chosen for  $\theta$  is determined by (16.5.3), regardless of the distribution of  $\theta$ . The value of  $\pi_g$  is determined by the relation of the frequency distribution of  $\theta$  to the item characteristic curve. Almost all examinees in group  $B$  have better than a 50% chance of answering the item correctly; therefore  $\pi_g$ , the proportion of correct answers, will be much higher than 0.5 in group  $B$ . In group  $A$ , on the other hand, most examinees will fail the item, so here  $\pi_g$  will be less than 0.5.

Consider the practical problem of building, over a period of years, a large pool of items to be used for constructing parallel test forms. Suppose, for example, that we are to pretest a set of newly written items on a group of examinees that differs somewhat from earlier pretest groups. To learn about any differences between this pretest group and previous pretest groups, we should administer a substantial number of "calibration" items, for which item parameters are already available from previous pretests, along with the newly written items.

Next we plot the values of  $a_g$  for the calibration items estimated from the new group ("group 2") against the values of  $a_g$  estimated from a previous "standard" pretest group ("group 1"). Because of the invariance of  $a_g$ , these two determinations would be identical except for the single fact that in each pretest group the origin and unit of measurement for  $\theta$  are determined so that  $\mu_\theta = 0$  and  $\sigma_\theta = 1$ . Thus the two sets of  $a_g$  ought to have a linear relationship,

with a slope equal to  $\sigma_1(\theta)/\sigma_2(\theta)$  and an intercept of zero. We may estimate this linear relationship by plotting the group-2  $a_g$  against the group-1  $a_g$  for the calibration items. We can then use the line so determined to convert group-2 values of  $a_g$  for the newly written items into "standard" values of  $a_g$  comparable to those that would have been obtained if the items had been pretested in group 1, the "standard" group.

The same procedure may be carried out with the  $b_g$ , except that the linear relationship so obtained does not in general have an intercept of zero. The key point is that under the normal ogive model, the  $b_g$  from the two groups have a strict linear relationship, whereas this is not true of the usual "item difficulty"  $\pi_g$ , nor of the inverse-normal transformation of  $\pi_g$  denoted by  $\gamma_g$  in (16.9.3). [The  $\gamma_g$  are linearly related to the College Entrance Examination Board "delta", which is described in Gulliksen (1950, pp. 368-369) and Conrad (1948)]. Since  $b_g = \gamma_g/\rho'_g$ , by (16.8.5), it is clear that if two sets of  $b_g$  are linearly related, then the corresponding  $\gamma_g$  cannot be linearly related unless all items have the same  $\rho'_g$ .

The conclusion is that if the normal ogive model holds, then the parameters  $a_g$  and  $b_g$  can be used to build a file of items with known characteristics. The parameters  $\pi_g$ ,  $\gamma_g$ , and  $\rho'^2_g$  cannot in general be so used unless all pretesting is done on one group, or unless all pretest groups are comparable to each other.

We can use Eqs. (16.9.3) and (16.10.7) to compute the parameters  $a_g$  and  $b_g$  from  $\pi_g$  and  $\rho'_{bg}$  only if  $\theta$  is  $N(0, 1)$  (if  $\theta$  is normally distributed, its mean can be taken as origin and its standard deviation can be taken as unit of measurement). In practical work, we never know the distribution of  $\theta$ , nor is its form easy to estimate from the data at hand. Good practical methods for estimating  $a_g$  and  $b_g$ ,  $g = 1, 2, \dots, n$ , without knowing the distribution of  $\theta$ , are not easy to find. The method outlined by Haley (1952) may be of interest for this purpose. In Section 17.9, a successful method for estimating  $a_g$  and  $b_g$  without knowledge of the frequency distribution of  $\theta$  will be outlined.

Can we expect the normal ogive model to hold to a good approximation for actual test data? We shall consider this question under four distinct sub-headings.

**Item homogeneity.** First, is a one-dimensional latent space an adequate practical approximation for actual test items? It is most likely to be so for tests that appear as though they ought to be homogeneous; for example, certain tests of vocabulary, reading, spelling, and some kinds of spatial ability. On the other hand, we should expect a mathematics test made up half of arithmetic reasoning items and half of plane geometry items to show at least  $k = 2$  latent dimensions.

A number of workers have factor-analyzed tetrachoric item intercorrelations for various tests to see if there is more than one common factor. They have often found that the residuals after extraction of one factor are near the size that one would expect from sampling fluctuations. Indow and Samejima (1966, pp. 24-25) administered the 30-item *LIS Measurement Scale for Non-Verbal*

*Reasoning Factor* to 883 students. They extracted the first nine latent roots of the matrix of tetrachoric item intercorrelations, using estimated communalities in the diagonal. Figure 16.11.2 shows how the size of the latent roots dropped off after the first was extracted. Since the smallest latent roots are treated as "noise", for practical purposes, and since the second latent root is almost as small as the later ones, there seems good reason to treat their data as arising from a one-dimensional latent space.

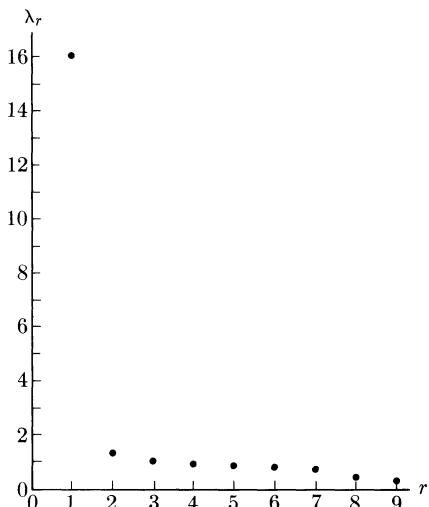


FIG. 16.11.2. The nine largest latent roots  $\lambda_r$  in order of size, for the correlation matrix of LIS items.

Actually, that the tetrachoric item intercorrelations have just one common factor is a sufficient but not a necessary condition for the unidimensionality of the latent space. The tetrachorics cannot be expected to have just one common factor except under certain normality assumptions, whereas such distributional considerations are irrelevant for the dimensionality of the complete latent space. Furthermore, if the tetrachorics have just one common factor, then the interitem phi coefficients will in general have more than one common factor, as the reader can readily verify from almost any numerical example where the  $n > 2$  items are of unequal difficulty. Therefore, *the number of common factors in a correlation (or covariance) matrix depends on the type of correlation coefficient (or covariance) used. It also depends on how the item scores are transformed before the correlations are computed. The dimensionality of the complete latent space does not depend on distributional assumptions, nor on a choice of a measure of interitem correlation, nor on any choice of transformation of the latent variables. Thus the dimensionality of the complete latent space is a more basic concept than is the number of common factors.*

The problem of statistically investigating the hypothesis of unidimensionality without specifying the shape of the item characteristic curves has not been completely solved. Some investigations of this hypothesis have been reported by Kirkham (1960) and Levine (1967).

**Normal ogive characteristic curves.** Second, given that the complete latent space is one-dimensional, does the assumption of normal ogive characteristic curves provide an adequate practical approximation for actual test items?

The utility of the normal ogive model (and of other similar models) can in principle be investigated by the following four steps:

1. Estimate the parameters of the model, assuming it is true (see Section 17.9).
2. Predict various observable results from the model, using the estimated parameters.
3. Consider whether the discrepancies between predicted results and actual results are small enough for the model to be useful ("effectively valid") for whatever practical application the investigator has in mind.
4. If in step 3 the discrepancies were considered too large, then it may be useful to compare them with the discrepancies to be expected from sampling fluctuations.

Step 3 is the crucial one for determining the utility of the model for use with samples of the size under investigation. If in step 3 we find the discrepancies to be less than perfectly satisfactory, the fault may lie not with the model but with the size of sample used. We shall then need step 4 to decide whether further research should be directed toward changing the model, or toward a further investigation of the same model using a larger sample of data.

It is important to note that neither in physics nor in psychology does one ever prove that a model is "really true". One merely shows that it is or is not in reasonable agreement with observed data. The possibility always remains that additional data will prove the model to be false. In fact, at least with psychological models of the type considered here, it can be taken for granted that every model is false and that we can prove it so, if only we collect a sufficiently large sample of data. The key question, then, is the practical utility of the model, and not its ultimate truthfulness.

Lord (1952) and Indow and Samejima (1962, 1966), using the assumption that  $\theta$  was normally distributed, found that the normal ogive model predicted univariate and bivariate observed-score distributions for their data rather well.

**Guessing.** The studies by Lord and by Indow and Samejima referred to above are both based on free-response items, where there was little opportunity for guessing. Third, then, does the normal ogive model apply to multiple-choice and other item types where there is much opportunity for guessing?

If examinees who do not know the answer to a multiple-choice item choose at random among the  $A$  alternative responses (an unrealistic assumption, as we have said before, made here for illustrative purposes), then the relative frequency of correct answers for such examinees will be  $1/A$ . In this case,  $P_\theta(\theta)$  should be asymptotic to the value  $1/A$  for low values of  $\theta$ . Thus the normal ogive model cannot hold for tests where the correct answer can be obtained by guessing.

The normal ogive model can be modified to allow for guessing. We shall not treat this modified model here, however, since a three-parameter logistic model that allows for guessing is considered in detail in Chapters 17 through 20.

**Speeded tests.** Almost all group tests are speeded for some examinees, since the test administrator seldom wishes to allow enough time for every single person to finish. If an examinee does not have time to read an item, then his response (if any) does not depend on his  $\theta$ . Therefore, fourth, the item characteristic curve model does not apply if examinees fail to finish the test in the time allowed.

In pure speed tests, the items would be extremely easy if the examinee could proceed at his leisure. For such tests, the Poisson or gamma models discussed in Chapter 21 should be applicable. For the usual time-limit test containing difficult as well as easy items, a combination of the Poisson or gamma models and the item characteristic curve approach needs to be worked out.

We shall not give further applications of the normal ogive model here since more powerful results for the logistic and three-parameter logistic models will be presented in Chapters 19 and 20.

## 16.12 Conditional Distribution of Test Scores

This section and the following ones deal primarily with general results that apply to any set of item characteristic functions  $P_g(\boldsymbol{\theta})$  for dichotomous items  $g = 1, 2, \dots, n$ . It is *not* assumed that these functions are normal ogives, nor even that the latent space is one-dimensional. We make two basic assumptions for the remainder of the present chapter:

**Assumption 16.12.1.**  $P_g \equiv P_g(\boldsymbol{\theta})$  is a continuous monotonic increasing function of each  $\theta_r$  for  $-\infty < \theta_r < \infty$ ,  $r = 1, 2, \dots, k$ , for any fixed set of values for the remaining  $\theta$ .

**Assumption 16.12.2.** If  $\theta_1, \theta_2, \dots, \theta_k$  are all sufficiently large, then  $P_g(\boldsymbol{\theta})$  will be arbitrarily close to one.

These assumptions seem reasonable for aptitude and achievement test items.

In principle, at least, an examinee with *sufficiently* high ability should have a probability near one of answering any given item correctly. Because of the possibility of guessing the answer to an item, it is *not* assumed here that  $P_g(\boldsymbol{\theta})$  approaches zero for low values of  $\theta_1, \theta_2, \dots, \theta_k$ .

If the test score  $X$  is the number of right answers, so that

$$X_* \equiv \sum_{g=1}^n U_{g*}, \quad (16.12.1)$$

then, from (16.3.7), the conditional distribution of  $X$  for given  $\boldsymbol{\theta}$  is the compound binomial distribution

$$f(x | \boldsymbol{\theta}) = \sum_{\sum u_g=x} \left( \prod_{g=1}^n P_g^{u_g} Q_g^{1-u_g} \right), \quad (16.12.2)$$

where the large summation sign is over all patterns of response  $(u_1, u_2, \dots, u_n)$  containing exactly  $x$  correct answers. This distribution is the same (Kendall and Stuart, 1958, Section 5.10) as that of the number of white balls obtained when one ball is drawn at random from each of  $n$  urns, the probability of drawing a white ball from urn  $g$  being  $P_g$ ,  $g = 1, 2, \dots, n$ . The quantity on the right-hand side of (16.12.2) is the coefficient of  $t^x$  in the expansion of the probability generating function

$$\prod_{g=1}^n (Q_g + P_g t).$$

*In the simple case of equivalent items, where all  $P_g(\theta)$  are identical functions of  $\theta$  for all  $g$ , the conditional distribution of  $x$  for given  $\theta$  is the ordinary binomial distribution*

$$f(x | \theta) = \binom{n}{x} P_g^x Q_g^{n-x}.$$

The mean of a compound binomial distribution is known to be  $\sum_1^n P_g$ . This fact provides the following basic theorem, which the reader should compare with Theorem 16.4.1.

**Theorem 16.12.1.** *The regression of test score on ability is proportional to  $\bar{P}(\theta)$ , the average of the item characteristic curves:*

$$\mu_{X|\theta} = \sum_{g=1}^n P_g(\theta) \equiv n\bar{P}(\theta). \quad (16.12.3)$$

The conditional variance of the test score is

$$\sigma_{X|\theta}^2 = \sum_{g=1}^n P_g Q_g. \quad (16.12.4)$$

In the terminology of conventional test theory, this important quantity is the squared standard error of measurement at a given ability level. This error variance is always less than the binomial variance  $n\bar{P}\bar{Q}$  except where all items are equivalent, that is, except where  $P_g(\theta) \equiv \bar{P}(\theta)$  for all items.

It follows from Assumption 16.12.2 and from (16.12.4) that the standard error of measurement is close to zero for examinees with sufficiently high ability. A small standard error of measurement is normally associated with "good" measurement. Thus it might seem paradoxical that a test has the smallest standard error of measurement for those examinees for whom the test is too difficult to measure effectively. The fact to be noted is that the effectiveness of the test as a measuring instrument at different ability levels depends not only on the standard error of measurement at those levels, but on other things as well. One possible index of the discriminating power of the test at different ability levels is the test information curve developed in Chapter 20.

### 16.13 A Relation of Latent Trait to True Score

What is the relation of latent trait  $\theta$  to true score? First let us rewrite Eqs. (16.12.3) and (16.12.4) to give the conditional mean and variance of the proportion-correct score  $Z = X/n$ :

$$\mu_{Z|\theta} = \bar{P}, \quad (16.13.1)$$

$$\sigma^2_{Z|\theta} = \frac{1}{n^2} \sum_{g=1}^n P_g Q_g. \quad (16.13.2)$$

The regression (16.13.1) of  $Z$  on  $\theta$ , of fundamental importance in latent trait theory, is called the *test characteristic curve*. It is denoted by  $\bar{P}(\theta)$ , or just by  $\bar{P}$ . Figure 16.13.1 shows the test characteristic curve for the nine-item test whose item characteristic curves are shown in Fig. 16.5.2.

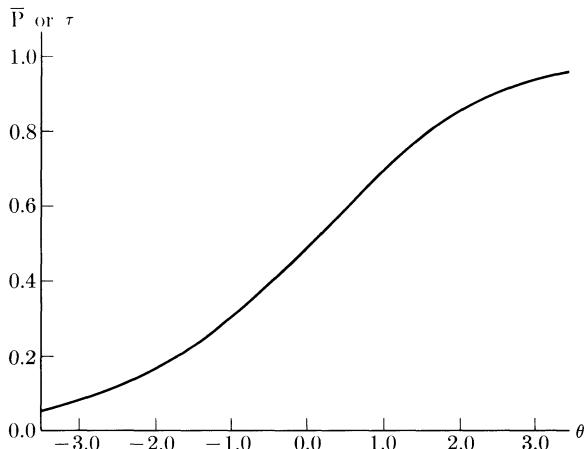


FIG. 16.13.1. Test characteristic curve for items shown in Fig. 16.5.2: The relation of true score to  $\theta$ .

Consider now a set of test forms that are “parallel” to the given test in the sense that the  $g$ th item in every form has the same characteristic function  $g = 1, 2, \dots, n$ . Let  $T_a$  now be the expected *proportion* of items that examinee  $a$  answers correctly over an infinite number of such forms. Thus  $T_a = \lim_{n \rightarrow \infty} z_a$  is the specific (relative) true score of examinee  $a$ .

In (16.13.2),  $\sigma^2_{Z|\theta} \rightarrow 0$  whenever  $n \rightarrow \infty$ . Thus the joint distribution of  $\theta$  and  $T$  is degenerate with no scatter about the regression of  $T$  on  $\theta$ . In other words, *the specific true score  $T$  has a functional rather than a statistical relation to the latent traits  $\theta_1, \theta_2, \dots, \theta_k$* .

*If the complete latent space is one-dimensional, the latent trait  $\theta$  is the same as the true score  $T$ , except for the scale of measurement used to describe it. Thus  $T$  and  $\theta$  are related by some monotonic increasing transformation.* If the complete latent space is multidimensional, then  $T$  is a single-valued function of  $\theta$ , but  $\theta_1, \theta_2, \dots, \theta_k$  are not uniquely determined by  $T$ .

If an  $n$ -item test is increased to infinite length by adding parallel tests,  $\bar{P}$  remains unchanged as  $n$  increases. Since all the frequency falls on the regression curve  $\mu_{T|\theta} = \bar{P}$ , we have

**Theorem 16.13.1.** *The functional relation between true score and latent traits is given by the test characteristic curve:*

$$T = \bar{P}(\theta) \equiv \frac{1}{n} \sum_{g=1}^n P_g(\theta). \quad (16.13.3)$$

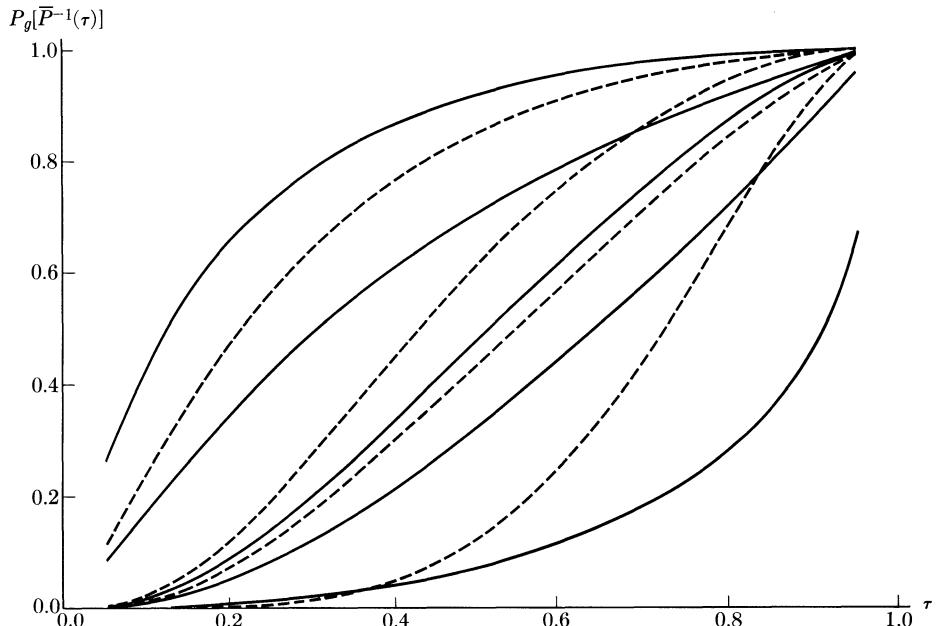


FIG. 16.13.2. Regression of item score on true score  $\tau$  for items shown in Fig. 16.5.2.

Figure 16.13.2 shows the regression of item score on true score for the nine items shown in Fig. 16.5.2. These last curves are obtained simply by applying the nonlinear transformation (16.13.3) to the base line of Fig. 16.5.2. Note that the item true-score regressions in Fig. 16.13.2 are not normal ogives. In fact, Theorem 16.4.1 applies: If the curves are averaged at each value of  $\tau$ , then the average curve will be a  $45^\circ$  straight line.

## 16.14 Typical Distortions in Mental Measurement

In this section, we assume that  $\theta$  is one-dimensional—we say that the test at hand “measures” just one trait  $\theta$ . Whereas all tests that “measure the same latent trait” by definition have the same  $\theta$ , each test usually has its own peculiar true-score metric. The test characteristic curve (tcc) shows how true score on each test is related to  $\theta$ . We can understand the peculiarities of each test as a

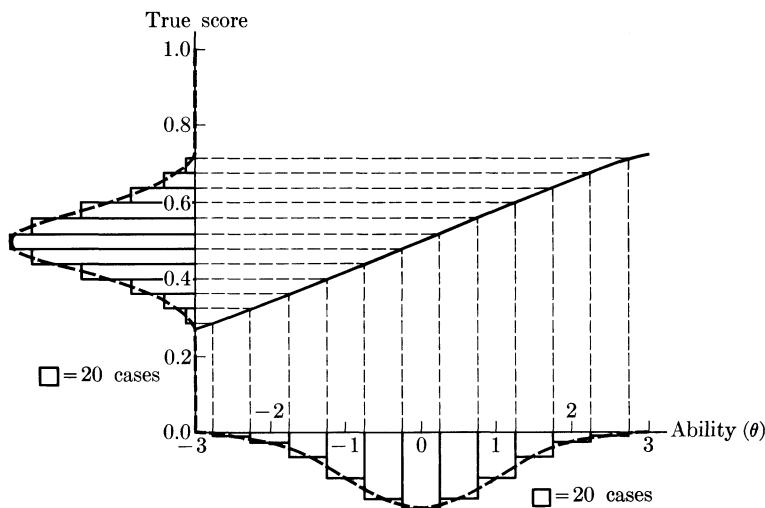


FIG. 16.14.1. Derivation of the distribution of true scores for a test of moderate difficulty and poor discriminating power.

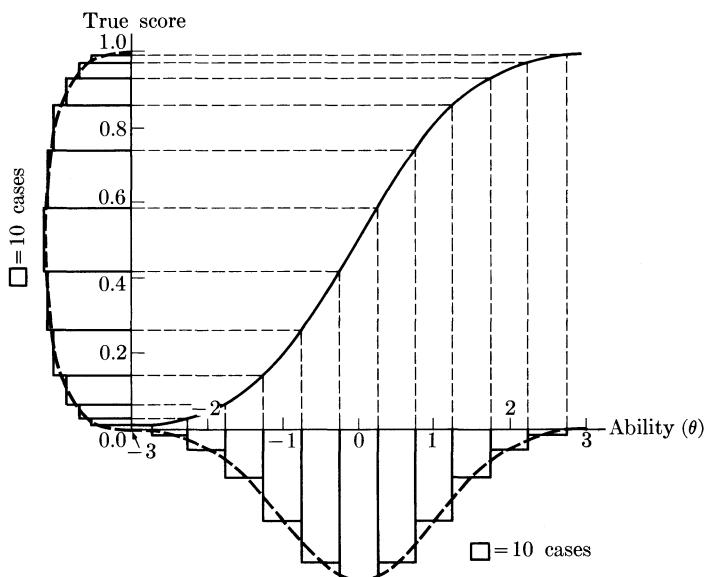


FIG. 16.14.2. Derivation of the distribution of true scores for a test of moderate difficulty and high discriminating power.

[Note: Figures 16.14.1 through 16.14.6 are taken from F. M. Lord, The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953, 13, 517-549. Used by permission.]

measuring instrument by considering its tcc. Since all the frequency in the bivariate distribution of  $T$  and  $\theta$  falls along the tcc, one can use the tcc to infer from any given distribution of  $\theta$  the corresponding distribution of  $T$ .

Figure 16.14.1 shows an almost straight tcc with low slope. This kind of curve characterizes a test that is composed of poorly discriminating items, or a test that has a very wide range of item difficulty. As we can see from the figure, the lower the slope of the tcc, the less the variance of the true-score distribution. If the tcc is almost straight throughout the range of  $\theta$  in the group tested,  $T$  is a nearly linear transformation of  $\theta$ ; hence the distribution of  $T$  will have nearly the same shape as that of  $\theta$ .

If a tcc were a strictly straight sloping line, the line would intersect the  $\theta$ -axis; thus examinees with sufficiently low  $\theta$  would have a negative  $T$ . Since  $T$ , being a proportion, cannot be negative, it follows from this contradiction that the tcc cannot be a straight line, except in the uninteresting case where it is horizontal and all examinees have the same true score.

The more discriminating the test items and the less the range of item difficulty, the steeper the tcc and the sharper its curvature. Figure 16.14.2 shows a typical free-response test with high discriminating power. The implications of the curvature of the tcc are plain from a visual comparison of the frequency distribution of the latent trait with that of the true score. Examinees who are near the extremes of the distribution of  $\theta$  get squeezed together in the true-score distribution. Thus if  $\theta$  happens to be normally distributed, the distribution of true score will typically be platykurtic. This is hardly surprising in view of the unrestricted range of  $\theta$  and the restricted range of the true score.

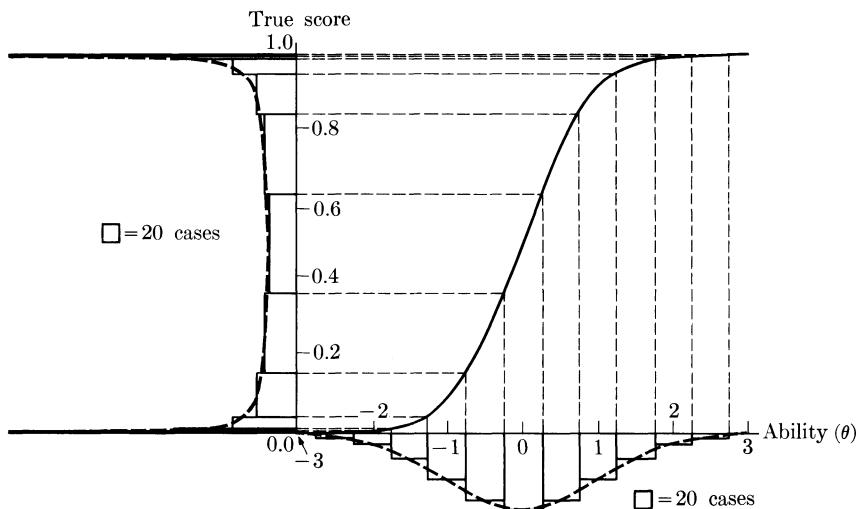


FIG. 16.14.3. Derivation of the distribution of true scores for a test of moderate difficulty and extremely high discriminating power.

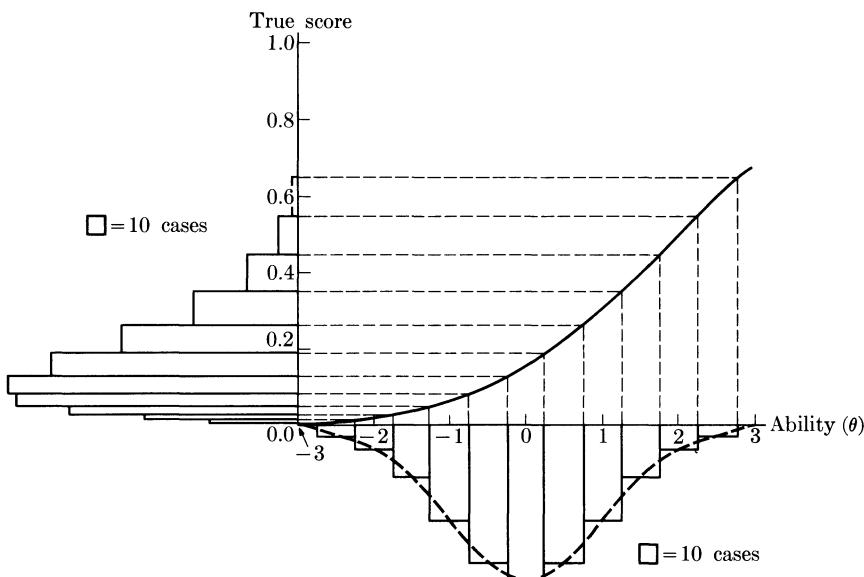


FIG. 16.14.4. Derivation of the distribution of true scores for a difficult test of average discriminating power.

Figure 16.14.3 shows what would happen if a sufficiently discriminating test could be built. The same picture also results when an ordinary test is administered to an extremely heterogeneous group of examinees. (In these figures, the unit of measurement of  $\theta$  is arbitrarily fixed so that  $\sigma_\theta = 1$  in the group tested; consequently it is impossible to tell from Fig. 16.14.3 whether it is the test that is extraordinarily discriminating, or the group that is exceptionally heterogeneous; the biserial correlation  $\rho'_{\theta g}$  for the individual test items can be high from either cause.) The U-shaped true-score distribution shown in the figure is seldom found in aptitude and achievement testing. It would be a very desirable true-score distribution to have whenever the purpose of the test is to select the top half of the examinees. The reason for this is that fewer examinees are near the cutting point if the distribution is U-shaped than if the distribution is bell-shaped. On the other hand, the U-shaped distribution is not a desirable one to have when the purpose of the test is to award one or two scholarships. It is visually clear from the figure that the test is discriminating poorly among examinees with very high  $\theta$ .

Figure 16.14.4 shows how a difficult test tends to produce a skewed distribution of scores, even when  $\theta$  is symmetrically distributed. Figure 16.14.5 shows how a test could produce a leptokurtic score distribution (in practice, leptokurtic distributions are rare). Figure 16.14.6 shows how the combination of an easy and a difficult subtest can produce a tcc like the one in the preceding figure; the middle curve is simply the average of the tcc's of the two half-tests.

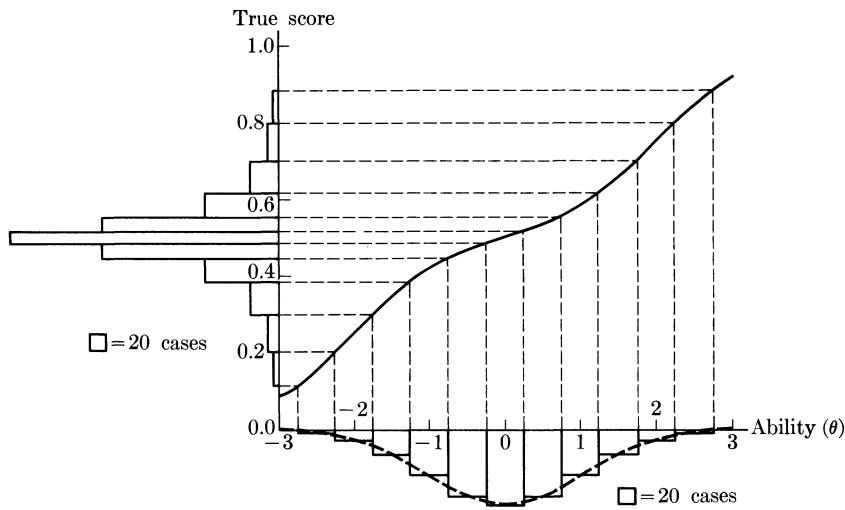


FIG. 16.14.5. Derivation of the distribution of true scores for a test composed half of easy items and half of difficult items.

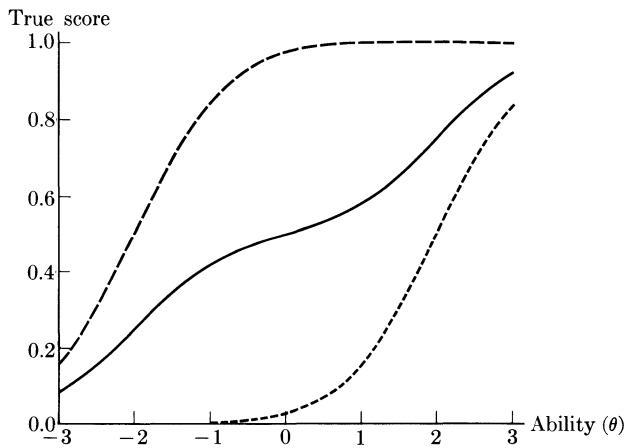


FIG. 16.14.6. Characteristic curves of an easy subtest (---), a difficult subtest (-----), and the combined test (—).

Figures 16.14.1 through 16.14.5 show true-score distributions arising from various tcc's. For a sufficiently long test, the (relative) observed-score distribution is much the same as the (relative) true-score distribution. The reader can imagine the nature of the observed-score distribution for a shorter test by making allowances for random errors of measurement.

Some conclusions for test construction are obvious from the foregoing discussion. If the examiner wants the observed-score distribution to have the same shape as the distribution of  $\theta$  [as does Brogden (1957)], then his test characteristic curve must be nearly straight throughout the relevant range of  $\theta$ . After trying out various combinations of item characteristic curves, the reader will find that such a test characteristic curve can be obtained either by using relatively undiscriminating items (low  $a_g$ ) or by using a very wide range of item difficulties ( $b_g$ ).

If the examiner wants to spread his observed-score distribution out as much as possible, he needs a test characteristic curve that is as steep as possible near the middle. Such a curve is obtained when the items are each as discriminating as possible (high  $a_g$ ) and are all of about medium difficulty ( $b_g$  near zero) for the group to be tested.

If the examiner wants to use the test only to select the top 2.5% of the examinees, then he wants to spread out the observed scores near the top of the score range. If ability is normally distributed in the group, he needs a test characteristic curve that is as steep as possible for  $\theta = 1.96$ . Such a curve is obtained by using difficult items with high  $a_g$ . Moreover all items should be of the same difficulty, namely,  $b_g = 1.96$ .

If an examiner wants to discriminate well in both tails of the distribution of ability, but does not need any discrimination among examinees of middling ability, then he might well build a test composed half of easy items and half of difficult items, as illustrated in Figs. 16.14.5 and 16.14.6. A more rigorous and detailed discussion on how to build optimal tests for specific purposes appears in Chapters 19 and 20.

### **References and Selected Readings**

- ANDERSON, T. W., Some scaling models and estimation procedures in the latent class model. In U. Grenander (Ed.), *Probability and statistics*. New York: Wiley, 1959, pp. 9-38.
- BOCK, R. D., Fitting a response model for  $n$  dichotomous items. Paper read at the Psychometric Society Meeting, Madison, Wisconsin, March 1967.
- BROGDEN, H. E., Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, **11**, 197-214.
- BROGDEN, H. E., New problems for old solutions. *Psychometrika*, 1957, **22**, 301-309.
- CONRAD, H. S., Characteristics and uses of item-analysis data. *Psychological Monographs: General and Applied*, 1948, **62**, No. 8 (Whole No. 295).
- CRONBACH, L. J., and H. AZUMA, Internal-consistency reliability formulas applied to randomly sampled single-factor tests: an empirical comparison. *Educational and Psychological Measurement*, 1962, **22**, 645-665.

- CRONBACH, L. J., and J. C. MERWIN, A model for studying the validity of multiple-choice items. *Educational and Psychological Measurement*, 1955, **15**, 337-352.
- CRONBACH, L. J., and W. G. WARRINGTON, Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika*, 1952, **17**, 127-147.
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950.
- HALEY, D. C., Estimation of the dosage mortality relationship when the dose is subject to error. Stanford: Applied Mathematics and Statistics Laboratory, Stanford University, Technical Report No. 15, 1952.
- INDOW, T., and FUMIKO SAMEJIMA, *LIS measurement scale for non-verbal reasoning ability*. Tokyo: Nihon-Bunka Kagakusha, 1962. (In Japanese.)
- INDOW, T., and FUMIKO SAMEJIMA, On the results obtained by the absolute scaling model and the Lord model in the field of intelligence. Yokohama: Psychological Laboratory, Hiyoshi Campus, Keio University, 1966. (In English.)
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*, Vol. I. New York: Hafner, 1958.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*, Vol. II. New York: Hafner, 1961.
- KIRKHAM, R. W., Scaling test scores from the law of comparative judgment. University of Western Australia, 1960, unpublished dissertation.
- LAWLEY, D. N., On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 1943, **61**, 273-287.
- LAWLEY, D. N., The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh*, 1944, **62-A**, 74-82.
- LAZARSFELD, P. F., Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science*, Vol. 3. New York: McGraw-Hill, 1959, pp. 476-542.
- LAZARSFELD, P. F., Latent structure analysis and test theory. In H. Gulliksen and S. Messick (Eds.), *Psychological scaling: theory and applications*. New York: Wiley, 1960, pp. 83-95.
- LAZARSFELD, P. F., The algebra of dichotomous systems. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961, pp. 111-157.
- LEVINE, M., Estimating item characteristic curves through uniform systems analysis. Paper read at the Psychometric Society Meeting, Madison, Wisconsin, March 1967.
- LORD, F. M., A theory of test scores. *Psychometric Monograph*, 1952, No. 7.
- LORD, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18**, 57-75. (a)
- LORD, F. M., The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953, **13**, 517-548. (b)
- MADANSKY, A., Determinantal methods in latent class analysis. *Psychometrika*, 1960, **25**, 183-198.
- MCDONALD, R. P., A general approach to nonlinear factor analysis. *Psychometrika*, 1962, **27**, 397-415.

- McDONALD, R. P., Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika*, 1967, **32**, 77-112.
- McDONALD, R. P., Nonlinear factor analysis. *Psychometric Monograph*, in press.
- MEREDITH, W., Some results based on a general stochastic model for mental tests. *Psychometrika*, 1965, **30**, 419-440.
- PATERSON, J. J., An evaluation of the sequential method of psychological testing. East Lansing: Office of Research and Publications, College of Education, Michigan State University, June 1962.
- TORGERSON, W. S., *Theory and methods of scaling*. New York: Wiley, 1958, pp. 360-402.
- TUCKER, L. R., Maximum validity of a test with equivalent items. *Psychometrika*, 1946, **11**, 1-13.

Part 5

**SOME LATENT TRAIT MODELS  
AND THEIR USE IN  
INFERRING AN EXAMINEE'S ABILITY**

*(Contributed by Allan Birnbaum)*



# SOME LATENT TRAIT MODELS

## 17.1 Introduction

In this chapter we shall consider in detail several models of tests, some of which have been introduced more briefly above (Sections 15.6 and 16.1 through 16.5). We shall now describe these models in self-contained mathematical terms to prepare ourselves to examine them, subsequently, in relation to theories and applications of tests. These models have been developed primarily in connection with tests of various general or special abilities, although it has proved of interest to consider them also in relation to the study of other kinds of traits, such as attitudes. For convenience, we shall refer to the trait in question simply as "ability".

We consider here tests consisting of items each to be scored 0 or 1, with  $u_g$  as the generic symbol for the score on item  $g$  and with  $\mathbf{v}' = (u_1, \dots, u_g, \dots, u_n)$  representing the set of scores, or the *response-pattern*, on a test of  $n$  items. This notation tacitly refers to scores of some one individual subject; when necessary, scores of a subject indexed  $a$  can be denoted more explicitly by  $\mathbf{v}'_a = (u_{1a}, \dots, u_{ga}, \dots, u_{na})$ .

Item scores  $u_g$  are related to an ability  $\theta$  by functions that give the probability of each possible score on an item for a randomly selected examinee of given ability. These functions are

$$Q_g(\theta) = \text{Prob} (U_g = 0 \mid \theta)$$

and the *item characteristic curve (ICC)*

$$P_g(\theta) = \text{Prob} (U_g = 1 \mid \theta) = 1 - Q_g(\theta).$$

These formulas are conveniently combined in the probability distribution function of  $U_g$ :

$$f_g(u_g \mid \theta) \equiv \text{Prob} (U_g = u_g \mid \theta) = P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \equiv \begin{cases} P_g(\theta) & \text{if } u_g = 1, \\ Q_g(\theta) & \text{if } u_g = 0, \end{cases}$$

where  $f_g$  is defined in a persons or in a persons-by-replications space.

We note that the regression function of any item response  $u_g$  is identical with its item characteristic curve since

$$\mathcal{E}(U_g | \theta) = 1 \cdot f_g(1 | \theta) + 0 \cdot f_g(0 | \theta) = P_g(\theta).$$

Any item for which  $P_g(\theta)$  has a constant value independent of  $\theta$  is not an indicant (and hence *a fortiori* not a measure) of  $\theta$  in the sense of Section 1.4. In most cases of interest here, we shall have  $P_g(\theta)$  strictly increasing in  $\theta$ , so that  $u_g$  will be an indicant and a measure of  $\theta$ . We do not assume a probability distribution for  $\theta$  in any part of the present treatment of this subject. For an extension of the theory which makes use of this assumption, the reader should see Birnbaum (1967).

These functions do not determine unequivocally the relation between an ability and a complete response pattern  $\mathbf{v}' = (u_1, \dots, u_n)$  unless they are supplemented in some definite way. The additional assumption found most useful in test theory and its applications, as well as the simplest assumption mathematically, is *local independence* (see Section 16.3). This assumption implies the mathematical condition of *statistical independence between responses* by a subject to different items; it is represented by the usual probability product form

$$\begin{aligned} \text{Prob} (\mathbf{V} = \mathbf{v} | \theta) &\equiv \text{Prob} (U_1 = u_1, \dots, U_n = u_n | \theta) \\ &= \text{Prob} (U_1 = u_1 | \theta) \cdots \text{Prob} (U_n = u_n | \theta) \\ &= \prod_{g=1}^n P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}. \end{aligned}$$

For example, the product form  $\text{Prob} [(U_1, U_2) = (1, 1) | \theta] = P_1(\theta)P_2(\theta)$  represents the fact that any subject of ability  $\theta$  gives independent responses to items 1 and 2; that is, that the probability  $P_2(\theta)$  of his correctly answering item 2 is the same as the conditional probability of his correctly answering item 2, given that he has correctly answered item 1. The relations of this assumption to more general models and theories, in which several abilities are considered jointly, have already been discussed in Section 16.2.

One basic aspect of the questions of validity and empirical and theoretical content discussed in Chapter 1, as they apply to the models introduced here, may be illustrated conveniently at this point. Consider any item, and consider a series of groups of subjects in which each subject is assumed to have common ability. Suppose that the probabilities of correct responses to the item in the respective groups are  $p_i$ , where  $0 \leq p_1 < p_2 < \dots < p_m \leq 1$ . Since we have mentioned all the empirically meaningful aspects of a model of a single item, we still remain free to choose arbitrarily a series of numbers  $\theta_i$ ,

$$-\infty < \theta_1 < \theta_2 < \dots < \theta_m < \infty,$$

which we may call the *true ability scores* of the respective groups. The choice

of these numbers  $\theta_i$  amounts to a choice of the specific form of an ICC function that shall represent the first item, since we *define* the function  $P_1(\theta)$  as the correspondence between respective ability scores  $\theta_i$  and values  $P_i = P_1(\theta_i)$ . Equivalently, given the numbers  $p_i$ , we can adopt *any* increasing function  $P_1(\theta)$  as the ICC of the item: This choice associates an ability score  $\theta_i$ , determined by  $p_i = P_1(\theta_i)$ , with the group of subjects scoring  $p_i$ .

These comments illustrate the fact that an essentially conventional element exists in the relations between ability levels  $\theta$  and observable item responses. Once any specific strictly increasing form has been adopted for  $P_1(\theta)$ , for example,  $P_1(\theta) = \Phi(2\theta - 1)$ , the statement that a subject has ability  $\theta = 2.1$  has empirical content and consequences in the contexts of models discussed here. For any second item (assuming local independence), the item characteristic curve  $P_2(\theta)$  has a value at  $\theta = 2.1$  which is estimable from empirical data in the same sense as is  $P_1(2.1)$ . Thus we are *not free* to adopt by definition *any* number as the value of  $P_2(2.1)$ . Similarly we are not free to adopt *any* assumption restricting even partially the possible functional forms of any other item characteristic curves  $P_g(\theta)$ ,  $g = 2, 3, \dots, n$ . This illustrates the fact that in general it is empirically meaningful (nontautological) to assume that any specific model, or even any class of models of partially restricted form, is valid in relation to a specified population of items. Therefore it is possibly false and hence is subject to empirical confirmation (or partial confirmation or disconfirmation). On the other hand, the assumption that any chosen *single* item has an item characteristic curve of a specified functional form  $P_g(\theta)$  that depends on ability  $\theta$  is, when considered *in isolation*, acceptable in principle as a definition of the ability scale of  $\theta$  values and is not an empirical specification.

## 17.2 The Logistic Test Model

A function which very nearly coincides with the normal ogive model treated in Section 16.5, and which has advantages of mathematical convenience in several areas of application, is the logistic (cumulative) distribution function

$$\Psi(x) = e^x / (1 + e^x) \equiv 1 / (1 + e^{-x}), \quad -\infty < x < \infty. \quad (17.2.1)$$

The inverse function is  $x = \log [\Psi/(1 - \Psi)]$ . For simple descriptive purposes, any graph of a cumulative normal distribution function  $\Phi(x)$  would serve equally well to illustrate this function, since it has been shown (Haley, 1952, p. 7) that

$$|\Phi(x) - \Psi[(1.7)x]| < 0.01 \quad \text{for all } x. \quad (17.2.2)$$

We may state this relation in another way: The logistic cdf  $\Psi(x)$  differs by less than 0.01, uniformly in  $x$ , from the normal cdf with mean zero and standard deviation 1.7; that is,

$$|\Phi(x/1.7) - \Psi(x)| < 0.01 \quad \text{for all } x.$$

The probability density function (pdf) corresponding to the logistic cdf is

$$\psi(x) = e^{-x}/(1 + e^{-x})^2 \equiv \Psi(x)[1 - \Psi(x)] \equiv \tanh^{-1}(x). \quad (17.2.3)$$

Berkson (1957) has given detailed tables of  $\Psi(x)$  and  $\psi(x)$ . Of course, tables of the exponential function and of the hyperbolic tangent are also available, and hence direct computation of values of these functions is not difficult.

The *logistic test model* is determined by assuming that item characteristic curves have the form of a logistic cumulative distribution function:

$$P_g(\theta) = \Psi[DL_g(\theta)] \equiv [1 + e^{-DL_g(\theta)}]^{-1} = [1 + e^{-D(a_g(\theta - b_g))}]^{-1}, \quad (17.2.4)$$

where  $L_g(\theta) = a_g(\theta - b_g)$ , and  $g = 1, 2, \dots, n$ . We have also

$$Q_g(\theta) = 1 - \Psi[DL_g(\theta)] \equiv [1 + e^{DL_g(\theta)}]^{-1},$$

$$P_g(\theta)/Q_g(\theta) = e^{DL_g(\theta)}, \quad \text{and} \quad \frac{\partial}{\partial \theta} P_g(\theta) = Da_g P_g(\theta) Q_g(\theta).$$

(Again, we do not interpret  $P_g(\theta)$  here as a probability distribution function, even when it has the mathematical properties of one.) Here  $a_g$  and  $b_g$  are item parameters whose roles are generally the same as those of the item parameters in the normal ogive model because of the qualitative, and nearly exact quantitative, similarity between the models. The symbol  $D$  denotes a number that serves, at our convenience, as a unit scaling factor. To maximize agreement between quantitative details in the normal and logistic models, we can and usually shall take  $D = 1.7$ ; then

$$P_g(\theta) = \Psi[1.7a_g(\theta - b_g)] \equiv (1 + e^{-1.7a_g(\theta - b_g)})^{-1}. \quad (17.2.4a)$$

For notational convenience, however, we shall often write the logistic model using the symbol  $D$  for the number 1.7.

We may view the logistic form for an item characteristic curve as a mathematically convenient, close approximation to the classical normal form, introduced to help solve or to avoid some mathematical or theoretical problems that arise with the normal model. Or we may view it as the form of a test model that is of equal intrinsic interest and of very similar mathematical form. The important questions of the validity of such models in observational and theoretical contexts are discussed elsewhere (see Sections 16.1 and 17.10).

The probability distribution function of a response  $u_g$  in a logistic test model is

$$\begin{aligned} f_g(u_g | \theta) &\equiv P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \\ &\equiv Q_g(\theta)[P_g(\theta)/Q_g(\theta)]^{u_g} \end{aligned} \quad (17.2.5)$$

$$= \frac{\exp [Da_g(\theta - b_g)u_g]}{1 + \exp [Da_g(\theta - b_g)]}, \quad (17.2.6)$$

and, under the assumption of local independence, the probability distribution function of a response pattern  $\mathbf{v}' = (u_1, \dots, u_n)$  is

$$\begin{aligned} \text{Prob}(\mathbf{V} = \mathbf{v} | \theta) &= \prod_{g=1}^n f_g(u_g | \theta) = \prod_{g=1}^n Q_g(\theta) \prod_{h=1}^n \exp [D a_h(\theta - b_h) u_h] \\ &= \left[ \prod_{g=1}^n Q_g(\theta) \right] \left[ \exp \left( \theta D \sum_{g=1}^n a_g u_g \right) \right] \left[ \exp \left( -D \sum_{g=1}^n a_g b_g u_g \right) \right]. \end{aligned} \quad (17.2.7)$$

The principal features of mathematical simplicity that characterize the logistic test model are, as we shall see, implicit in this last form. In particular, "all the information about  $\theta$  available in a response pattern  $\mathbf{v}$ " (in a sense to be specified) is given by the particular test score formula

$$x = x(\mathbf{v}) = \sum_{g=1}^n a_g u_g,$$

which does not depend on the difficulty parameters  $b_g$ . We may further illustrate the roles of item parameters and the properties of such a test score formula by considering an artificial example of the logistic test model. Let us take just four of the items whose parameters have the values represented in Fig. 16.5.1, namely,  $g = 3, 4, 5$ , and  $6$ . (The same figure serves equally well here to illustrate either logistic or normal item characteristic curves.) We have  $a_3 = 100$ ,  $a_4 = 100$ ,  $a_5 = 1$ , and  $a_6 = 1$ . The test score is then

$$\begin{aligned} x &= 100y_3 + 100y_4 + y_5 + y_6 \\ &= 100(y_3 + y_4) + (y_5 + y_6). \end{aligned}$$

The possible values of  $x$  are just

0	1	2
100	101	102
200	201	202.

We see that the major part of this ordering of subjects' response patterns, which is represented by the rows of the preceding array, is determined by the heavily weighted responses to the informative items  $y_3$  and  $y_4$ . The only role of the less informative items in this example is to give a finer ordering compatible with the initial rough ordering. This example is extreme: Typical tests one meets in practice have more items and less extreme variation in weights  $a_g$ . With more nearly typical tests, it is usually possible to reverse an ordering of two response patterns based only on responses to several items if all items are taken into account in a suitable weighted composite score.

**Table 17.3.1**  
Standard deviation of sample item-test biserials\*

Test	Number of items	Sample item- test biserials	
		Mean	Standard deviation
Listening Comprehension	50	0.51	0.12
English Structure	70	0.48	0.11
Vocabulary	60	0.55	0.09
Reading Comprehension	30	0.54	0.09
Writing Ability	60	0.44	0.11

\* From an internal Educational Testing Service report (SR-66-80) prepared by Dr. Frances Swineford.

### 17.3 Other Models

If we assume a common value for the discriminating powers of the items, each  $a_g = 1$ , say, and take  $D = 1$ , we obtain the form

$$P_g(\theta) = \Psi(\theta - b_g) \equiv (1 + e^{b_g - \theta})^{-1}.$$

We can write

$$\theta^* = e^\theta \quad \text{and} \quad b_g^* = e^{b_g}$$

to denote, respectively, an ability parameter and an item difficulty parameter, each represented on a transformed scale. Then we have

$$P_g(\theta) \equiv P_g^*(\theta^*) = \left(1 + \frac{b_g^*}{\theta^*}\right)^{-1} = \frac{\theta^*}{b_g^*} \left(1 + \frac{\theta^*}{b_g^*}\right)^{-1}.$$

Rasch (1960) has developed the test model of this restricted logistic form. We see that this model is a special case of the logistic model in which all items have the same discriminating powers, and all items can vary only in their difficulties. Whenever this special logistic model holds, the considerable body of theoretical and practical methods developed by Rasch is applicable (see Chapter 21).

One very important question emerges at this point: Do the items in a test really differ from each other in discriminating power? This question is crucial to evaluating the validity of the models and methods of this and the following three chapters and to comparing these evaluations with evaluations of the validity of the simpler models and methods of Chapter 21. Some available item analysis data suggest an affirmative answer for multiple-choice paper and pencil tests. These data, which are represented in Table 17.3.1, are based on a sample of 3805 examinees. The table shows the mean and standard deviation of the sample biserial correlation between item score and test score for each of

five different tests. If the true biserial correlation is 0.50 in a normal population of this size ( $N = 3805$ ), then the standard error of a biserial correlation from a sample of this size will only be from about 0.016 to about 0.019, depending on the item difficulty. (An approximate formula appears in McNemar, 1962, Eq. 12.3.) Since the standard deviations in this particular sample are at least five times as large as this standard error, it is clear that the variation found here among item-test biserials is almost entirely due to real differences among the item discriminating power parameters. In this sample we find that even if we disregard the five percent of the items with the highest and the five percent with the lowest discriminating power parameters, we still have a range from about 0.31 to about 0.67. Since item-test biserials approximate item-ability biserials, whose close relation to the slope of the item characteristic curve was discussed rather fully in Section 16.10, it is clear that the item characteristic curves of the items in Table 17.3.1 differ from each other by more than a mere translation (change of origin).

If  $\theta > b_g$ , then for any fixed values of  $b_g$  and  $\theta$ ,

$$\Phi[a_g(\theta - b_g)] \quad \text{and} \quad \Psi[D a_g(\theta - b_g)]$$

both increase to 1 as  $a_g$  increases; and if  $\theta < b_g$ , then both decrease to 0 as  $a_g$  increases. We may represent these limiting values formally as

$$\Phi[\infty(\theta - b_g)] = \Psi[\infty(\theta - b_g)] \equiv \begin{cases} 1 & \text{if } \theta > b_g, \\ 0 & \text{if } \theta < b_g, \end{cases}$$

since

$$(\theta - b_g)\infty \equiv \begin{cases} +\infty & \text{if } \theta > b_g, \\ -\infty & \text{if } \theta < b_g. \end{cases}$$

For convenience, we can give the value 1 to the otherwise undefined symbols  $\Phi(\infty \cdot 0)$ ,  $\Psi(\infty \cdot 0)$ . Then we may define an item characteristic curve by

$$P_g(\theta) = \Phi[\infty(\theta - b_g)] \quad \text{or} \quad \Psi[\infty(\theta - b_g)].$$

These may be considered extreme, limiting cases of ICCs within the normal ogive and the logistic test models. Such ICCs do not have the property, generally assumed above, of increasing continuously and strictly as  $\theta$  increases. Each is characterized fully by a single difficulty parameter  $b_g$ ; for abilities  $\theta < b_g$  it has the value zero, and at this ability level it increases discontinuously to unity. These curves may be regarded as representing items whose responses  $y_g$  are error-free indicants of abilities, in the sense that taking  $y_g = 1$  as indicating  $\theta \geq b_g$  and  $y_g = 0$  as indicating  $\theta < b_g$  entails probability zero of erroneous indications for each possible value of  $\theta$ . It may be said that ICCs of this extreme form give "perfect scaling", since an ordering of subjects' abilities  $\theta$  on the basis of any test consisting of such items is error-free (with probability 1, or certainty). Such items are basic to the scaling methods and the theory

developed by Guttman (1950), particularly in connection with scaling of latent traits  $\theta$  representing attitudes.

Lazarsfeld has developed several classes of latent trait models, but primarily for the investigation of attitudes rather than abilities. One of these may conveniently be described here. If

$$P_g(\theta) = a_g(\theta - b_g), \quad g = 1, \dots, n,$$

where  $\theta$  is restricted to an interval on which all values of  $P_g(\theta)$  lie between 0 and 1, we have the *linear model*. Despite quantitative differences, here, as in other models described above, the item parameter  $a_g$  represents discriminating power in the sense of rate of change of  $P_g(\theta)$  with respect to  $\theta$ , and  $b_g$  locates the part of the  $\theta$  scale where the item is effective. In Chapter 24, we shall present several other models developed by Lazarsfeld.

Methodological problems related to these models are discussed briefly by Torgerson (1958, Ch. 13), who gives references to basic papers and subsequent work. A later discussion is that of Lazarsfeld (1959).

Even subjects of very low ability will sometimes give correct responses to multiple-choice items, just by chance. One model for such items has been suggested by a highly schematized psychological hypothesis. This model assumes that if an examinee has ability  $\theta$ , then the probability that he will know the correct answer is given by a normal ogive function  $\Phi[a_g(\theta - b_g)]$  of exactly the kind considered in Section 16.5; it further assumes that if he does not know it he will guess, and, with probability  $c_g$ , will guess correctly. It follows from these assumptions that the probability of an incorrect response is

$$Q_g(\theta) = \{1 - \Phi[a_g(\theta - b_g)]\}(1 - c_g),$$

and that the probability of a correct response is the item characteristic curve

$$P_g(\theta) = c_g + (1 - c_g)\Phi[a_g(\theta - b_g)]. \quad (17.3.1)$$

The psychological hypothesis implicit here has been mentioned primarily to point up a mathematical feature of this form; the empirical validity of this form is not dependent on this psychological hypothesis. This model is possibly more reasonable than the random-guessing models discussed in Sections 14.3 and 14.5.

The function (17.3.1) approaches its minimum  $c_g$  as  $\theta$  decreases. Its graph is that of a normal ogive curve except that the range of ordinates 0 to 1 is replaced by the range  $c_g$  to 1. If one of five multiple-choice alternatives were chosen at random whenever guessing occurred, we would have  $c_g = \frac{1}{5}$ , as in Fig. 17.3.1, where the other item parameters are equal to those in Fig. 16.5.1. Each of the general illustrative comments above concerning the item parameters  $a_g$  and  $b_g$  of normal ogive models can be adapted to apply to their roles in these models.

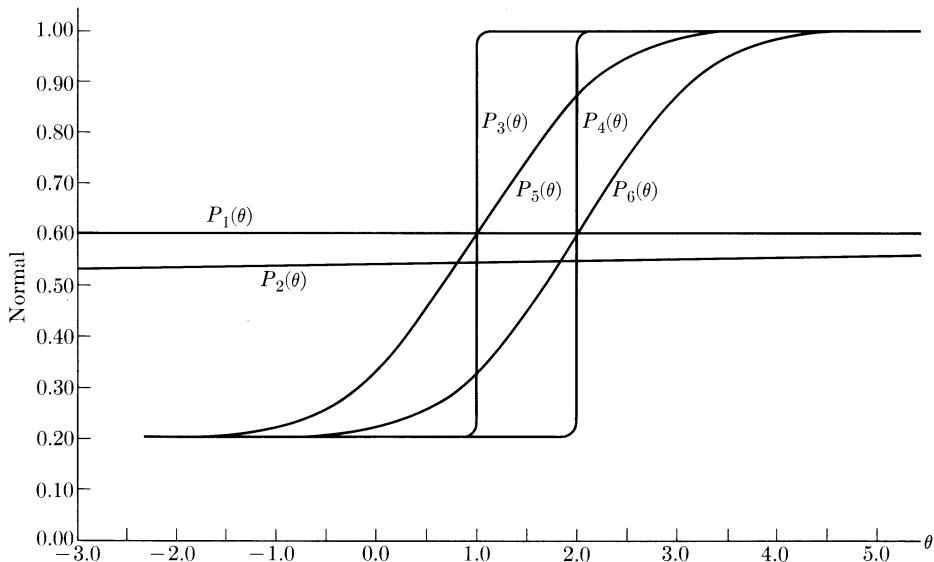


FIG. 17.3.1. Three-parameter normal ogive and logistic item characteristic curves.

Similarly with the logistic model, we may take account of guessing probabilities by using modified item characteristic curves, which here assume the form

$$P_g(\theta) = c_g + (1 - c_g)\Psi[D\alpha_g(\theta - b_g)],$$

which Fig. 17.3.1 serves to illustrate. More detailed consideration of the roles of item parameters in such models is given below.

#### 17.4 The Test as a Measuring Instrument: Examples of Classification and Estimation of Ability Levels by Use of Test Scores

We shall find it useful to consider the mathematical model of a test as having dual but related purposes. One purpose is to determine the value  $\theta$  of an examinee's ability with adequate precision; the second is to classify an examinee into ability categories with adequately small probabilities of misclassification. We shall present brief descriptions of some estimation and classification methods based on test scores. These will illustrate some of the applications of the theory that we shall develop. Each of the simplifying assumptions or restrictions made here will require critical reconsideration later.

We shall consider a model of a test, represented by a specified probability function

$$\text{Prob} [V' = (u_1, \dots, u_n) | \theta],$$

possibly having one of the forms described above, in which the ability  $\theta$  is the

only unknown parameter. We shall adopt a specified test score formula  $x = x(\mathbf{v}) \equiv x(u_1, \dots, u_n)$ . These two functions determine the cdf of the test score:

$$F(x | \theta) = \text{Prob} [X(\mathbf{V}) \leq x | \theta] \equiv \sum_{X(V) \leq x} \text{Prob} (\mathbf{V} = \mathbf{v} | \theta). \quad (17.4.1)$$

Numerical determinations of  $F(x | \theta)$  for a number of score formulas and tests will be illustrated. In the simplest case, that of items having identical characteristic curves

$$P \equiv P(\theta) \quad \text{and} \quad x = \sum_{g=1}^n u_g,$$

the cdf,  $F(x | \theta)$ , is just the binomial cdf for  $n$  trials with parameter  $P(\theta)$ :

$$F(x | \theta) = \sum_{k=0}^x \binom{n}{k} P^k Q^{n-k}, \quad x = 0, 1, \dots, n. \quad (17.4.2)$$

Local independence is assumed here.

In most cases of interest, the magnitudes of discontinuities in  $F(x | \theta)$  (that is, the probabilities of the individual possible values of  $x$ ) will all be small for each  $\theta$ , usually of the order of several percent or less. For many theoretical and practical purposes, it is convenient to treat  $F(x | \theta)$  as continuous in  $x$  for each fixed  $\theta$ , and also it is sometimes convenient to employ specific continuous functions of  $x$  as working approximations subject to appropriate bounds or independent checks on the approximations entailed. For illustrative simplicity, we treat  $F(x | \theta)$  in this section as continuous and assume that for each fixed  $\theta$ , it is strictly increasing from 0 to 1 with  $x$ . In the preceding binomial example, the convenient approximation is the usual one by the normal cdf (see, for example, Lindgren, 1962, p. 149):

$$F(x | \theta) \equiv \sum_{k=0}^x \binom{n}{k} P^k Q^{n-k} \doteq \Phi \left[ \frac{x + \frac{1}{2} - nP}{(nPQ)^{1/2}} \right], \quad (17.4.3)$$

which is continuous in  $x$  for each  $\theta$ .

We further assume throughout this section that for each fixed value of  $x$ ,  $F(x | \theta)$  is strictly and continuously decreasing from 1 to 0 with  $\theta$ ; the respective distributions of  $x$  are said to be *stochastically ordered* when this condition holds. In the binomial example, this condition holds for both the exact and approximate formulas for  $F(x | \theta)$ , given that  $x < k$ . This condition is entailed by weak assumptions which are usually satisfied, namely, that each  $P_g(\theta)$  increases strictly and continuously with  $\theta$ , and that  $x(u_1, \dots, u_n)$  is nondecreasing in each  $u_g$  and increasing in at least one of them. The latter conditions hold in all cases described above.

When these conditions hold, the respective cdf's of scores of a given test can be represented conveniently in the manner illustrated in the schematic graphs of Figs. 17.4.1 and 17.4.2. Figure 17.4.2 is a schematic representation

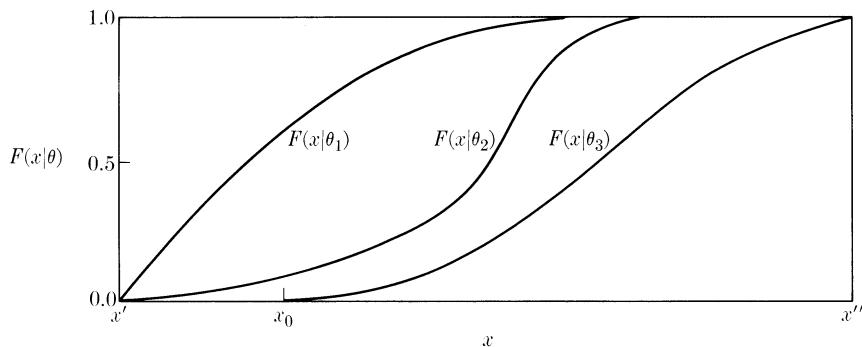


FIG. 17.4.1. Cdf's of scores  $x$  for several  $\theta$ -values,  $\theta_1 < \theta_2 < \theta_3$ .

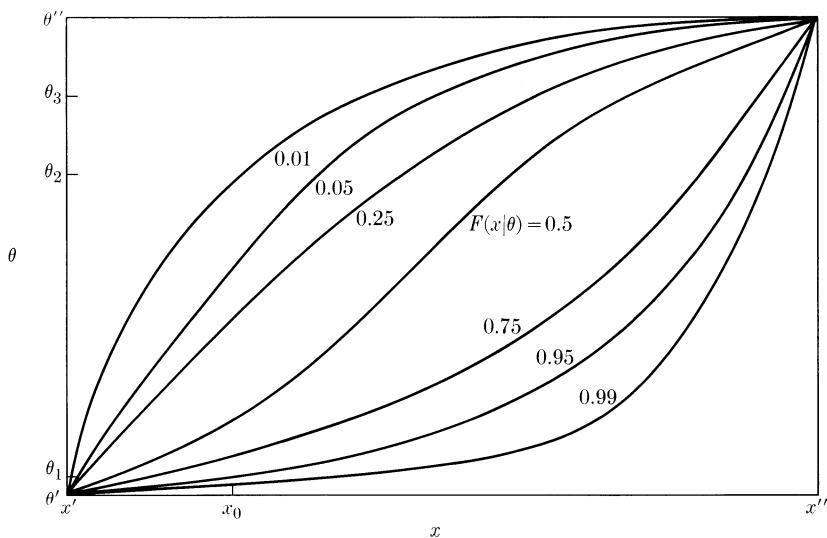


FIG. 17.4.2. Contours of constancy for cdf's  $F(x, \theta)$ .

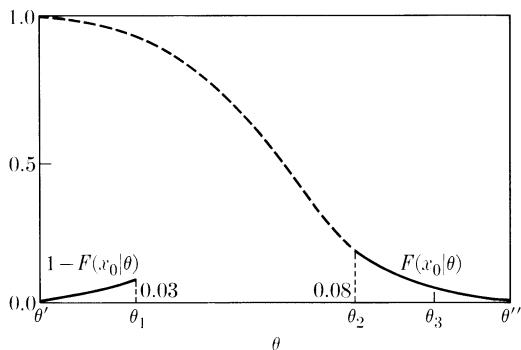


FIG. 17.4.3. Error probabilities of the classification rule that classifies high when  $x > x_0$ .

of the function of two arguments  $F(x | \theta)$  that map several “contours of constant height” of the  $F(x | \theta)$  “surface” over the  $(x, \theta)$  plane. Figure 17.4.1 represents three “sections” (“slices”) through this surface, made at  $\theta = \theta_1, \theta_2, \theta_3$ , respectively. Figure 17.4.3 represents in two forms one “section” made in the perpendicular direction at  $x = x_0$ ; we shall explain this figure below.

The discriminating power of a test is illustrated most simply in problems of discriminating between just two levels of ability. One common rule classifies those subjects whose scores exceed some specified number  $x_0$  as “high” and classifies others as “low”. With this rule, if  $\theta$  is any ability level considered definitely high, then  $F(x_0 | \theta) = \text{Prob}(X \leq x_0 | \theta)$  is the probability of erroneous (low) classification of a subject of that ability. Since  $F(x_0 | \theta)$  decreases as  $\theta$  increases, it is natural to focus attention on the smallest  $\theta$  value considered definitely high, say  $\theta_2$ . The rule’s maximum probability of erroneous classification of a high-ability subject is then  $F(x_0 | \theta_2)$ , as illustrated by Fig. 17.4.3. Similarly, if  $\theta_1$  is the highest ability considered definitely low, then

$$1 - F(x_0 | \theta_1) = \text{Prob}(X > x_0 | \theta)$$

is the rule’s maximum probability of erroneous classification of a low-ability subject. At abilities between  $\theta_1$  and  $\theta_2$ , neither classification is considered definitely erroneous and no error probabilities are considered.

By decreasing  $x_0$ , we can decrease  $F(x_0 | \theta_1)$ , the maximum misclassification probability for low abilities, but only at the cost of increasing  $1 - F(x_0 | \theta_2)$ , the maximum misclassification probability for high abilities. Evidently the possibility of circumventing such restrictions on the discriminating power attainable with a given test depends on basic reconsideration of the forms of test-score formulas and classification rules adopted; and these considerations might show that improvement requires the use of a different test.

We note that some of the present considerations parallel some of the interpretations given above of the discriminating power of single items in terms of item characteristic curves. The common element is the role of the rate of increase of  $P_g(\theta)$  and of  $1 - F(x_0 | \theta)$ , respectively, as  $\theta$  increases. So long as a test is used only to provide a classification rule based on a comparison of its scores with some fixed critical value  $x_0$ , the test is in effect equivalent to a single hypothetical test item having responses

$$u_1^* = \begin{cases} 1, & \text{corresponding to } x > x_0 \\ 0, & \text{corresponding to } x \leq x_0, \end{cases}$$

and item characteristic curve

$$P_1^*(\theta) = 1 - F(x_0 | \theta).$$

We can consider parameters describing the form of  $F(x_0 | \theta)$  and  $1 - F(x_0 | \theta)$  in rough analogy with the parameters of single items: For example, if  $F(x_0 | \theta') = \frac{1}{2}$ ,

then  $\theta'$  can be called the *difficulty level of the classification rule*; and

$$-\frac{\partial}{\partial \theta} F(x_0 | \theta)$$

evaluated at  $\theta = \theta'$  can be called the *discriminating power of the classification rule*. We shall consider in detail below the ways in which such parameters of the test and other properties of classification rules depend on the parameters of the respective test items. Parameters such as  $a_g$ ,  $b_g$  in logistic or normal items serve to characterize an item fully, and these parameters admit heuristically useful and relevant descriptive interpretations. However, their principal significance lies in their precise role in contributing to the information structure of a test, a notion we shall elaborate in the following sections and chapters. For a classification rule represented by a function  $1 - F(x_0 | \theta)$ , an analogous pair of parameters may be of some limited descriptive value, but in general they must fall far short of determining fully the course of  $1 - F(x_0 | \theta)$  and the values of all error probabilities of practical interest. A summary description of the error probabilities that is more useful for many purposes is a pair of points such as those represented in Fig. 17.4.3, which indicate that at the values  $\theta_1$ ,  $\theta_2$  the error probabilities of respective types are 0.03 and 0.08.

A standard technique of estimation, that of confidence limits, is directly applicable when the distributions of test scores are available graphically, as in Fig. 17.4.3, or equivalently in tables of percentage points. A lower confidence limit estimator with a confidence coefficient of 95%, say, is defined as any statistic  $t(\mathbf{v})$  having the property that

$$\text{Prob}[t(\mathbf{V}) \leq \theta | \theta] = 0.95 \quad \text{for each } \theta.$$

That is, for each possible value  $\theta$  of an examinee's ability, the probability is 0.95 that the estimate  $t(\mathbf{v})$  derived from the response pattern of such an examinee will be a correct lower bound on his ability.

In the case at hand, where  $\mathbf{v}$  is represented just by a test score  $x$ , it is easy to obtain a statistic  $t(x)$  with the above property. Let  $x^*$  denote the numerical test score of an examinee. Let  $\theta^*(x^*, 0.95)$  denote the number  $\theta^*$  that satisfies the equation  $F(x^* | \theta^*) = 0.95$ ; in Fig. 17.4.2,  $\theta^*$  corresponds to  $x^*$  in the sense that  $(x^*, \theta^*)$  is a point on the 0.95 contour. Then  $\theta^*$  is a lower 95% confidence limit estimate of the examinee's ability  $\theta$ . (The fact that  $\text{Prob}[\theta^*(X, 0.95) \leq \theta | \theta] = 0.95$  is an easily derived consequence of the definition of  $\theta^*$ .) Taking  $\theta^* = 1.3$ , for concreteness of illustration, we may record this conveniently in the notation:  $\text{Conf}(\theta \geq 1.3) = 0.95$ .

Other confidence limits are determined similarly. For example,  $\theta^*(X, 0.25)$  is an upper 75% confidence limit estimator, defined implicitly by  $F(x | \theta) = 0.25$  and having the basic property that

$$\text{Prob}[\theta^*(X, 0.25) > \theta | \theta] = 0.75 \quad \text{for each } \theta.$$

The pair of estimators,  $\theta^*(x, 0.95)$ ,  $\theta^*(x, 0.05)$ , together constitute a 90%

confidence interval estimator of  $\theta$ ; For each possible true value  $\theta$ , they include  $\theta$  between them with probability 90%. Among the various types of useful point estimators of  $\theta$ , one which we may conveniently describe here is  $\theta^*(x, 0.5)$ . This point estimator is median-unbiased, that is, it both overestimates and underestimates  $\theta$  with probability  $\frac{1}{2}$ .

The precision of a confidence interval estimator is represented by its confidence coefficient, together with the typical lengths of the interval estimates that it determines; or, more precisely and adequately, by error probabilities for over- or underestimation by various amounts. We shall indicate below how such precision properties of confidence intervals and confidence limits can be related in detail to the discriminatory power of a test in classification by ability levels.

### 17.5 The Information Structure of a Test and Transformations of Scale of Scores

When we apply a test model in conjunction with a specific test to such classification and estimation problems as the ones illustrated in the preceding section, we observe that no properties of the model play any role except the cdf's of the score of that specific test. For example, Fig. 17.4.2 might represent two different tests, each with very different numbers and types of items and item characteristic curves, but the estimation and classification methods based on scores of the respective tests would still have identical error-probability properties. This equivalence would hold even if the cdf's were different, but could be made to coincide when scores  $x$  of one test were transformed by a suitable increasing function  $x^*(x)$  into scores  $x^*$  of the second test. This is true because no properties of the scale of scores  $x$  beyond simple ordering have been used here. Thus, for such standard inference methods based on an adopted test score formula, we may consider the family of distributions of scores  $F(x | \theta)$  as representing the essential *information structure* or *canonical form* of a test, with the qualification that the scale of scores  $x$  plays only the role of simple ordering.

To illustrate this qualification, consider any given family of cdf's  $F(x | \theta)$ , any arbitrarily chosen ability  $\theta_2$ , and the function defined by

$$x^* \equiv x^*(x) = F(x | \theta_2).$$

This is a strictly increasing function of  $x$ , and we can adopt it to define scores  $x^*$  on a new scale; the range of such scores is  $0 \leq x^* \leq 1$ . Let the cdf's of such scores  $x^*$  be denoted by

$$F^*(x^* | \theta) \equiv \text{Prob}[X^*(\mathbf{V}) \leq x^* | \theta], \quad 0 \leq x^* \leq 1.$$

A special property of scores defined in this way is that when  $\theta = \theta_2$  [that is, the ability level that has been arbitrarily chosen for the definition of  $x^*(x)$ ], the distribution of scores  $X^*$  takes the special "uniform" form

$$F^*(x^* | \theta_2) \equiv x^*, \quad 0 \leq x^* \leq 1.$$

This property characterizes the *probability integral transformation*  $x^*(x)$ . An illustration appears in Fig. 17.5.1, a figure that is a transformed version of Fig. 17.4.1. Since such transformations of scores are typically nonlinear, expected values and variances of the transformed scores  $x^*$  do not have any simple relations to expected values and variances of scores  $x$  on the original scale. Thus the concepts and methods presented in this section are not closely linked with any use of moments of distributions of test scores at given ability levels.

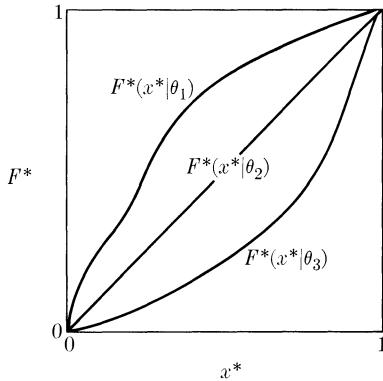


FIG. 17.5.1. Transformed version of Fig. 17.4.1.

[If the curve for  $\theta_3$  were deleted from Fig. 17.5.1 and the resulting figure were rotated, the new figure would be familiar to many students of mathematical statistics (see, for example, Lindgren, 1962, p. 236). For each  $\alpha$ , the test of the hypothesis  $H_0: \theta = \theta_2$  against the hypothesis  $H_1: \theta = \theta_1$ , based on rejecting  $H_0$  just when  $x^*$  is sufficiently small, is the test that rejects just when  $x^* \leq x_\alpha^* \equiv \alpha$ . The power of this test is given by  $1 - \beta = F^*(x_\alpha^* | \theta_1) \equiv F^*(\alpha | \theta_1)$ . Thus  $\beta = 1 - F^*(\alpha | \theta_1)$  gives the “ $\alpha, \beta$  curve”.]

## 17.6 Transformations of Scales of Ability

It is interesting to consider the preceding point concerning the scaling of *scores*, in combination with the point concerning the scaling of *abilities* illustrated at the end of Section 17.1, where a certain freedom in specification of the ability scale was discussed. The latter point can be applied here: Abilities  $\theta$  can be replaced by abilities  $\theta^* = \theta^*(\theta)$  on a transformed scale in such a way that the family of cdf's of scores

$$F^{**}(x^* | \theta^*) = \text{Prob}[X^*(V) \leq x^* | \theta^*]$$

is given any chosen form compatible with the other conditions thus far assumed. For example, the transformation  $\theta^*(\theta)$  defining the new scale of abilities can be chosen so that each possible score value  $x^*$  is the median of the distribution

of scores for ability level  $\theta^* = \theta^*(\theta) = x^*$ ; that is, so that

$$F^{**}(x^* | \theta^*) = \frac{1}{2} \quad \text{whenever} \quad x^* = \theta^* = \theta^*(\theta),$$

as in Fig. 17.6.1, which is a transformed version of Fig. 17.4.2. To prove this, we note that the condition  $F(x | \theta) = 0.5$  defines implicitly the function  $x(\theta)$ , the median of scores  $x$  for each ability  $\theta$ , in terms of the given cdf's. Hence  $x^*[x(\theta)]$  is the median of transformed scores  $x^*$  for each ability  $\theta$ . We are now free to define a transformation of abilities by

$$\theta^*(\theta) = x^*[x(\theta)].$$

Now for each ability  $\theta$ , the transformed ability  $\theta^* = \theta^*(\theta)$  coincides with the median of the distribution of transformed scores.

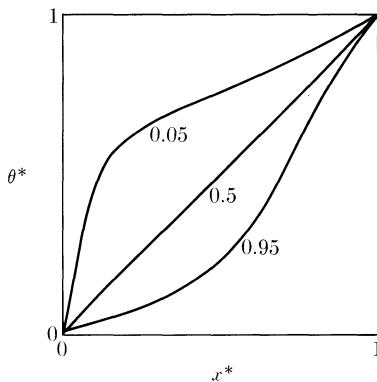


FIG. 17.6.1. Transformed version of Fig. 17.4.2.

We may mention another significant example of possible rescaling of scores and abilities. Let  $x^*(x)$  be any arbitrarily chosen strictly increasing function, subject only to the mild restriction that the expected values of scores,  $\mathcal{E}(X^* | \theta)$ , exist for each  $\theta$ . Let us determine a new scale of scores by the transformation  $x^*(x)$ . Next, we can choose a transformed scale of abilities  $\theta^*$ , determined by the transformation function  $\theta^*(\theta) = \mathcal{E}(X^* | \theta)$ . From the assumption that the cdf's  $F(x | \theta)$  are stochastically ordered, it follows that  $\theta^*(\theta)$  is an increasing function, and that the cdf's  $F^{**}(x^* | \theta^*)$  will also be stochastically ordered. Every scale of abilities  $\theta^*$  that may be determined in this way satisfies the essential condition for the definition of true score presented in Chapter 2, namely,  $\mathcal{E}(X^* | \theta^*) = \theta^*$  for each  $\theta^*$  (see Chapter 24).

It is interesting to consider an analogous question: If any test model and score formula are given and are represented by a specified family of cdf's  $F(x | \theta)$  having the two monotonicity properties assumed above, then is it always possible to keep the *given ability scaling* (which, of course, may have been obtained by an arbitrary transformation from a previous ability scaling), and also to realize simultaneously, by means of some monotone transformation  $x^*(x)$  of

the score scale, the essential condition for true-score theory, namely,  $\theta = \mathcal{E}(X^* | \theta)$  for each  $\theta$ ? The answer is, usually, no—the possibility depends on the detailed structure of the given cdf's  $F(x | \theta)$ . To illustrate this simply, we shall assume that  $x$  has a finite number of possible values

$$x_1 < \cdots < x_j < \cdots < x_M,$$

and consider an arbitrary sequence of different possible values of  $\theta$ , namely,  $\theta_1, \theta_2, \dots, \theta_i, \dots$ . Let  $C_{ij} = \text{Prob}(X = x_j | \theta_i)$  for each  $i, j$ . (Here we drop the assumption of continuity of the cdf's  $F(x | \theta)$ , an assumption which is typically inexact although useful elsewhere.) If  $x^*(x)$  is any monotone transformation, we may write

$$x_j^* = x^*(x_j) \quad \text{and} \quad x_1^* < \cdots < x_j^* < \cdots x_M^*.$$

If the transformed scores  $x^*$  are to satisfy the true score assumption

$$\theta = \mathcal{E}(X^* | \theta) \equiv \mathcal{E}[x^*(X) | \theta], \quad \text{for each } \theta,$$

then for each  $i$  we must have

$$\theta_i = \mathcal{E}[x^*(X) | \theta_i] \quad \text{or} \quad \theta_i = \sum_{j=1}^M C_{ij} x_j^*.$$

In general, such linear equations in  $M$  unknowns  $x_j^*$  are inconsistent, even when only  $M + 1$  such equations (determined by any chosen  $M + 1$  values  $\theta_i$ ) are considered in isolation. Thus the possibility of realizing the conditions for true score theory *for the given ability scale*, even by monotone transformation of the given score formula, is limited by and dependent on the detailed structure of the given model  $F(x | \theta)$ . This contrasts with the possibility of realizing the true score assumptions *for the given score scale*, which, as we have seen above, is always possible if a monotone transformation of the ability scale is allowed. The discussion here amplifies and formalizes the discussion in Chapter 2 of the relationship among various concepts of true score.

On the other hand, to explore the approximate applicability of classical true-score theory to a given model when the given ability scaling is to be retained, we can first choose successively values  $\theta_i$  that seem to represent the range of abilities of interest effectively and can then consider the sequence of equations

$$\theta_i = \sum_{j=1}^M C_{ij} x_j^*,$$

continuing so long as the equations are consistent and allow ordered solutions  $x_j^*$ . If any set of such equations does not determine unique, ordered solutions, we may supplement it by adding arbitrary and possibly convenient independent

linear restrictions on the  $x_j^*$ , possibly including specification of convenient values for

$$x_1^*, \quad x_M^*, \quad \frac{x_1^* + x_M^*}{2}, \quad \text{or} \quad \frac{1}{M} \sum_{j=1}^M x_j^*,$$

or some combination of these, until we have obtained  $M$  linearly independent equations.

Whenever  $F(x | \theta)$  is a normal cdf for each  $\theta$ , we may take  $\theta^*(\theta) = \mathcal{E}(X | \theta)$ , which is both the mean and the median of  $X$ , for each  $\theta$ . When  $F(x | \theta)$  is at least approximately a normal cdf, then  $\theta^*(\theta) = \mathcal{E}(X | \theta)$  is usually approximately the median (as well as exactly the mean) of  $X$ .

Of course, weak true-score theory is characterized by its use of no restrictive assumptions on the forms of the cdf's  $F(x | \theta)$  of scores other than low-order moments of scores. The preceding considerations illustrate some of the many connections and differences to be found between weak and strong true-score theories.

### 17.7 Calculations of Distributions of Test Scores

Applications of the inference methods illustrated above require adequate numerical determinations of the distributions of test scores at respective ability levels. In most practical work with cognitive tests, response patterns are represented only by test scores having the particular form

$$x = x(\mathbf{v}) = \sum_{g=1}^n w_g u_g \tag{17.7.1}$$

of weighted sums of item responses, where the  $w_g$  are specified numerical weights. Most commonly, the weights are specified as equal, either as  $w_g \equiv 1$ , where calculation of  $x$  then gives the number of correct responses, or as  $w_g \equiv 1/n$ , where calculation of  $x$  then gives the proportion of correct responses. In the following chapters, we shall see that in important cases a suitably chosen linear (or weighted-sum) score formula can be used to provide estimators with optimal or nearly optimal precision and classification rules of good discriminating power. In this section, we shall present some useful theoretical and computational methods for calculating distributions of test scores of this form and illustrate these by numerical examples of the applications illustrated above.

The principal result that we shall present here is the normal approximation to the cdf  $F(x | \theta)$  for score formulas  $x$  of any weighted sum form  $x = \sum_{g=1}^n w_g u_g$ , where the  $w_g$  are given constants. The theoretical basis for such normal approximations in the general case consists of the central limit theorems available for sums of nonidentical independent random variables (see, for example, Lindgren, 1962, p. 147, and Loève, 1955, p. 288). The resulting approximation formulas for  $F(x | \theta)$  depend on the given test model only through the mean

and variance of  $X$  for each  $\theta$ :

$$\mathcal{E}(X | \theta) = \sum_{g=1}^n \mathcal{E}(w_g U_g | \theta) \equiv \sum_{g=1}^n w_g P_g(\theta), \quad (17.7.2)$$

$$\sigma^2(X | \theta) = \sum_{g=1}^n \sigma^2(w_g U_g | \theta) \equiv \sum_{g=1}^n w_g^2 P_g(\theta) Q_g(\theta). \quad (17.7.3)$$

Then the approximation formula is

$$F(x | \theta) \doteq \Phi\{[x - \mathcal{E}(X | \theta)]/\sigma(X | \theta)\}. \quad (17.7.4)$$

In connection with various specific test models and problems of application below, the preceding general formulas for the moments of scores will be specialized and substituted in the last relation.

For test models with nonequivalent items, and for composite scores with unequal weights, we require here a form of the central limit theorem that allows nonidentically distributed terms  $w_g U_g$ . On the other hand, for many practical purposes we may conveniently interpret the hypothetical concept of increase without limit of the number  $n$  of nonequivalent test items as the case of a test model with  $G_n = ng$  items specified as follows: The first  $n$  items may have any specified ICCs; each successive set may consist of  $n$  items equivalent, respectively, to those of the first set; and  $G$  may increase without limit. The simplest case of the central limit theorem, that of identically distributed terms, applies here, since each set of  $n$  items can formally be considered to contribute a single term

$$Z_r = \sum_{g=1}^n w_g U_{nr+g}, \quad r = 0, 1, 2, \dots, \quad \text{to} \quad X = \sum_{r=0}^G Z_r,$$

provided that  $w_{nr+g} = w_g$  for  $r = 1, 2, \dots$ .

#### Examples: Moments and quantiles of test scores for items of various types.

##### *Moments of item responses*

$$\mathcal{E}(U_g | \theta) = P_g(\theta), \quad \sigma^2(U_g | \theta) = P_g(\theta)Q_g(\theta),$$

1. Normal ogive

$$\mathcal{E}(U_g | \theta) = \Phi[L_g(\theta)], \quad \sigma^2(U_g | \theta) = \Phi[L_g(\theta)]\Phi[-L_g(\theta)],$$

$$\text{where } L_g(\theta) = a_g\theta - b_g.$$

2. Logistic

$$\mathcal{E}(U_g | \theta) = \Psi[DL_g(\theta)], \quad \sigma^2(U_g | \theta) = \psi[DL_g(\theta)],$$

where

$$\psi(t) = \frac{\partial}{\partial t} \Psi(t) \equiv \frac{e^t}{(1 + e^t)^2}.$$

### 3. Three-parameter logistic

$$\begin{aligned}\mathcal{E}(U_g | \theta) &= c_g + (1 - c_g)\Psi[DL_g(\theta)] = \Psi[DL_g(\theta)] + c_g\Psi[-DL_g(\theta)], \\ \sigma^2(U_g | \theta) &= (1 - c_g)\psi[DL_g(\theta)] + c_g(1 - c_g)\Psi[-DL_g(\theta)]^2.\end{aligned}$$

*Moments of terms in locally best composite scores (developed below in Section 19.3)*

$$\begin{aligned}w_g(\theta) &= P'_g(\theta)/P_g(\theta)Q_g(\theta), \\ \mathcal{E}[w_g(\theta)U_g | \theta] &= [P'_g(\theta)/P_g(\theta)Q_g(\theta)]P_g(\theta) = P'_g(\theta)/Q_g(\theta), \\ \sigma^2[w_g(\theta)U_g | \theta] &= w_g(\theta)^2\sigma^2(U_g | \theta) = [P'_g(\theta)^2/P_g(\theta)^2Q_g(\theta)^2]P_g(\theta)Q_g(\theta) \\ &= P'_g(\theta)^2/P_g(\theta)Q_g(\theta).\end{aligned}$$

#### 1. Normal ogive

$$\begin{aligned}w_g(\theta) &= a_g\varphi[L_g(\theta)]/\Phi[L_g(\theta)]\Phi[-L_g(\theta)], \\ \mathcal{E}[w_g(\theta)U_g | \theta] &= \varphi[L_g(\theta)]/\Phi[-L_g(\theta)], \\ \sigma^2[w_g(\theta)U_g | \theta] &= a_g^2\varphi[L_g(\theta)]^2/\Phi[L_g(\theta)]\Phi[-L_g(\theta)].\end{aligned}$$

#### 2. Logistic

$$\begin{aligned}w_g(\theta) &= Da_g \quad (\text{uniformly best weights}), \\ \mathcal{E}(w_g U_g | \theta) &= Da_g\psi[DL_g(\theta)], \quad \sigma^2(w_g U_g | \theta) = D^2a_g^2\psi[DL_g(\theta)].\end{aligned}$$

#### 3. Three-parameter logistic

$$\begin{aligned}w_g(\theta) &= Da_g\Psi[DL_g(\theta) - \log c_g], \quad \mathcal{E}[w_g(\theta)U_g | \theta] = Da_g\Psi[DL_g(\theta)], \\ \text{Var } [w_g(\theta)U_g | \theta] &= (1 - c_g)D^2a_g^2\psi[DL_g(\theta) - \log c_g]\psi[DL_g(\theta)].\end{aligned}$$

*Moments of composite scores.* Dividing each weight  $w_g$  in a scoring formula by the same positive constant (for example, the sum of the weights) does not change the ratio between respective weights, which is the essential feature of the scoring formula. Therefore we may express any composite score formula in the form

$$x = \frac{\sum_{g=1}^n w_g u_g}{\sum_{g=1}^n w_g}.$$

For example, the weights  $w_g$  may be

- 1) equal weights, for instance,  $w_g = 1$  for  $g = 1, \dots, k$ ;
- 2) best weights (developed below in Section 18.4)

$$w_g(\theta_1, \theta_2) = \log \frac{P_g(\theta_2)Q_g(\theta_1)}{P_g(\theta_1)Q_g(\theta_2)};$$

or

- 3) locally best weights  $w_g(\theta) = P'_g(\theta)/P_g(\theta)Q_g(\theta)$  as developed in Section 19.3.  
Thus we may write the moments of any composite score  $x$  as, say,

$$\mathcal{E}(X | \theta) = \frac{\sum_{g=1}^n w_g P_g(\theta)}{\sum_{g=1}^n w_g} \quad \text{and} \quad \sigma^2(X | \theta) = \frac{\sum_{g=1}^n w_g^2 P_g(\theta) Q_g(\theta)}{\left( \sum_{g=1}^n w_g \right)^2} \equiv \sigma^2(\theta).$$

*Quantiles of composite scores under the normal approximation: a measure of information.* Using our previous assumption that a given composite score  $x$  has cdf's  $F(x | \theta)$  that are continuously strictly increasing in  $x$  and decreasing in  $\theta$ , we may implicitly define the  $(1 - \alpha)$ -quantile of  $X$ , which we denote by

$$x^*(1 - \alpha, \theta),$$

as the solution  $x$  of

$$F(x | \theta) = 1 - \alpha. \quad (17.7.5)$$

If we assume in particular a normal form for the cdf's  $F(x | \theta)$ , we have

$$x^*(1 - \alpha, \theta)$$

defined as the solution  $x$  of

$$F(x | \theta) \equiv \Phi\{[x - \mu(\theta)]/\sigma(\theta)\} = 1 - \alpha. \quad (17.7.6)$$

Taking  $\Phi^{-1}$  of both sides and solving for  $x$ , we then have

$$x^*(1 - \alpha, \theta) = \mu(\theta) + \Phi^{-1}(1 - \alpha)\sigma(\theta). \quad (17.7.7)$$

[The quantity  $\Phi^{-1}(1 - \alpha)$  is a normal deviate cutting off a normal-curve left-tail area of  $1 - \alpha$ .]

The composite score will actually approach a normal form with increasing  $n$ , under the slight restriction that the values  $w_g^2 P_g(\theta) Q_g(\theta)$  are uniformly bounded away from zero for the given  $\theta$ -value considered. (This follows from the central limit theorem for the case of nonidentically distributed terms; see, for example, Loève, 1955, p. 310.) Under mild additional conditions (which will often be satisfied, and which can be checked with reference to specific applications), formula (17.7.7) can be approximated adequately closely, over any interval of  $\theta$ -values centered at any given value  $\theta'$  and appreciably wide, by a linear function of  $\theta$ . This function of  $\theta$  may be written

$$x^*(1 - \alpha, \theta) = A + B(\theta - \theta'), \quad (17.7.8)$$

where

$$A = \mu(\theta') + \Phi^{-1}(1 - \alpha)\sigma(\theta') \quad \text{and} \quad B = \mu'(\theta'), \quad \mu' = \frac{\partial}{\partial \theta} \mu(\theta).$$

[This function represents the Taylor series approximation to (17.7.7),

$$\begin{aligned} x^*(1 - \alpha, \theta) &\doteq \mu(\theta') + \Phi^{-1}(1 - \alpha)\sigma(\theta') + \mu'(\theta')(\theta - \theta') \\ &\quad + \Phi^{-1}(1 - \alpha)\sigma'(\theta')(\theta - \theta'), \end{aligned}$$

further simplified by deleting the last term. This term may be deleted because

$$\sigma'(\theta') = \frac{\partial}{\partial \theta} \sigma(\theta)|_{\theta=\theta'}$$

tends to be negligible in comparison with  $\mu'(\theta')$ .] By solving (17.7.8) for  $\theta$ , we obtain the corresponding linear approximation

$$\theta^*(x, 1 - \alpha) = \theta' + [x - \mu(\theta') - \Phi^{-1}(1 - \alpha)\sigma(\theta')]/\mu'(\theta'). \quad (17.7.9)$$

Now the latter formula represents (approximately) the lower  $(1 - \alpha)$ -level confidence limit estimate of the ability  $\theta$  of an individual with score  $x$ , as discussed in Section 17.4 above. One natural and convenient indication of the value of a given test and scoring formula is the width of the resulting confidence interval estimates of ability. The width of the approximate  $(1 - 2\alpha)$ -level confidence interval indicated by the approximation (17.7.9) for any given  $\alpha < \frac{1}{2}$  is just  $\theta^*(\alpha, x) - \theta^*(1 - \alpha, x)$  as determined from (17.7.9):

$$[\Phi^{-1}(1 - \alpha) - \Phi^{-1}(\alpha)] \frac{\sigma(\theta')}{\mu'(\theta')}.$$

We see that this width is proportional to  $\sigma(\theta')/\mu'(\theta')$ , a constant independent of  $\alpha$ ,  $x$ , and  $\theta$  under the assumed approximation. For  $\theta$  near  $\theta'$ , therefore, this constant serves as an index of precision of interval estimation based on the given test and scoring formula. As it turns out, the same constant also characterizes the effectiveness of the test and scoring formula for a wide variety of other purposes. Hence we shall use the term *information* to designate the related quantity

$$I(\theta', x) = \mu'(\theta')^2/\sigma^2(\theta'). \quad (17.7.10)$$

More precisely, we shall refer to  $I(\theta', x)$  as the *information* provided by the given test and composite scoring formula in the neighborhood of  $\theta'$ . The function  $I(\theta, x)$  is called the *information function of the scoring formula*  $x$ . It should be noted that the symbol  $x$  appears here not as a variable argument of  $I(\theta, x)$ , but as an abbreviation for "the probability distributions  $F(x, \theta)$  of the scoring formula  $x$ ", in terms of which  $I(\theta, x)$  is defined. The definition is of course made with reference to some specified mental test, in terms of which the scoring formula is defined; thus, for a given scoring formula,  $I(\theta, x)$  is a function of  $\theta$  only.

There are two additional reasons for using the term “information” for this quantity:

1. Let us consider the error probability functions of classification rules based on  $x$ . At  $\theta'$ , the slope of each of these functions has a given value  $\alpha = \alpha(\theta')$ . With increasing  $n$ , these values tend to be proportional to

$$\sqrt{I(\theta', x)} \equiv \mu'(\theta')/\sigma(\theta').$$

*Proof.* Writing  $F(x | \theta) = \Phi\{[x - \mu(\theta)]/\sigma(\theta)\} = \Phi(t)$ , where  $t = [x - \mu(\theta)]/\sigma(\theta)$ , we see that the slope of the error probability function  $1 - F(x | \theta)$  is  $-\partial F(x | \theta)/\partial \theta$ . By the chain rule, this can be written as  $-[d\Phi(t)/dt](\partial t/\partial \theta)$ , or  $-\varphi(t)(\partial t/\partial \theta)$ . Hence

$$\begin{aligned}\frac{\partial t}{\partial \theta} &= -\frac{\{\sigma(\theta)\mu'(\theta) + [x - \mu(\theta)]\sigma'(\theta)\}}{\sigma^2(\theta)} = -\frac{\mu'(\theta)}{\sigma(\theta)} - t \frac{\sigma'(\theta)}{\sigma(\theta)} \\ &\doteq -\sqrt{I(\theta, x)},\end{aligned}$$

since, as we noted in the derivation of (17.7.9),  $\sigma'(\theta)$  is small compared with  $\mu'(\theta)$ .

2. With increasing  $n$ , when  $\theta'$  is the true value, the point estimator  $\theta^*(x, 0.5)$  tends to be normally distributed, with mean  $\theta'$  and variance  $1/I(\theta', x)$ .

Thus  $I(\theta, x)$  plays the role of an index of precision of estimation. If we are dealing with nonlinear scoring functions  $x = x(v)$ , then we cannot apply the central limit theorem in the direct way indicated in connection with (17.7.4) above. Nevertheless, for an important and wide class of nonlinear scoring functions and estimators, we can show that there is an approach to a limiting normal distribution with increasing  $n$ . The definition  $I(\theta, x)$  or  $I(\theta, \theta^*) = \mu'(\theta)^2/\sigma^2(\theta)$  is extended to such cases of nonlinear  $x$  or  $\theta^*$  by taking  $\mu(\theta)$  and  $\sigma^2(\theta)$  to represent the *asymptotic moments* of  $x$  and  $\theta^*$ . These asymptotic moments are moments of the limiting normal distributions, which are in theory, and in relevant examples, distinct from the limits of exact moments of  $x$  or  $\theta^*$ .

In particular, in Section 20.3, we shall consider the maximum likelihood estimator  $\hat{\theta}$  and its information function  $I(\theta, \hat{\theta})$  in some detail. As in the preceding special case, we shall see that the role of an index of precision of estimation is played quite frequently by the information function  $I(\theta, x)$ , for a given scoring formula, and  $I(\theta, \theta^*)$ , for a given test and estimator.\*

---

\* For derivations and discussions of these properties, see Cramér (1946, pp. 498–506) or Birnbaum (1961a, pp. 122–127). In such discussions of asymptotic distributions in connection with maximum likelihood, the results (1) and (2) above are obtained by replacing the “score”  $S(x, \theta) = (\partial/\partial \theta) \log f(x | \theta)$  by  $[x - \mathcal{E}(X | \theta)]/\sigma(X | \theta)$ .

These and other uses and interpretations of the information functions  $I(\theta, x)$  of various test models and composite score formulas will appear below, particularly in Chapter 20, where self-contained discussions of some aspects of information functions are given.

### 17.8 Quantal Response Models in General

The test models introduced in this chapter have analogues in other technical and scientific areas. Models of the general form  $\text{Prob}(\mathbf{V} = \mathbf{v} | \boldsymbol{\theta})$  have been called *quantal response models*. The normal ogive model (including the three-parameter case described above) has been used extensively in biological assay work. (See, for example, Finney, 1944 and 1952. In the second reference, comparisons between biological and psychometric applications are given.) The use of the logistic model as an alternative to the normal in bioassay work has also been developed extensively (see Berkson, 1953 and 1957). For another type of biological assay, the *dilution series* model with  $P_g(\theta) = 1 - e^{-a_g\theta}$  has been used (Fisher, 1922, pp. 363–366, and Cochran, 1950). Applications of such models have also been made in industrial gauging (Stevens, 1948) and genetics (for example, Rao, 1965, pp. 302–309, and Kempthorne, 1957, p. 181, and references therein). An appreciable part of the discussion in the next chapters has general relevance to quantal response models.

### 17.9 Estimation of Item Parameters

Two maximum likelihood methods have been given for estimating the item parameters in the normal ogive test model, by Tucker (1951) and by Lord (1953). These are discussed by Torgersen (1958, pp. 388–391), where they are related to other mathematical problems that arise in scaling. In the following paragraphs (1) and (2), we present two adaptations of these methods to the case of the logistic model. [For the restricted case of the logistic model described in Section 17.2, in which only the item difficulty parameters  $b_g$  are unknown, Rasch (1960) has given advantageous estimation methods. Many details of the derivation and calculation of estimates presented in the next paragraphs have forms similar to those of the more restricted estimation problem discussed in more detail in Section 20.3, which deals with maximum likelihood estimates of ability.]

The likelihood function of the responses observed when an  $n$ -item test is administered to a group of  $N$  examinees of abilities  $\theta_1, \theta_2, \dots, \theta_N$  is

$$L = \prod_{c=1}^N \prod_{g=1}^n \{1 - \Psi[Da_g(\theta_c - b_g)]\} \exp[Da_g(\theta_c - b_g)u_{gc}]. \quad (17.9.1)$$

Let

$$x_c = \sum_{g=1}^n u_{gc}$$

denote the raw score of examinee  $c$ . Then

$$\mathcal{E}(X_c \mid \theta_c) = \sum_{g=1}^n \Psi[D a_g(\theta_c - b_g)]$$

is an increasing function of  $\theta_c$ , provided that all  $a_g$  are positive. For two examinees of abilities  $\theta_c$  and  $\theta_{c'} > \theta_c$ , we have  $\text{Prob}\{X_{c'} > X_c\} \rightarrow 1$  as  $n$  increases, provided that the  $a_g$  are bounded away from zero and the  $b_g$  are bounded. That is, there is a tendency for ability order to be reflected correctly in the ordering of raw scores, as the number of items increases.

1. If we assume that the examinees are a random sample from a population in which the ability  $\theta$  has a standard normal (or logistic) distribution, then, as  $N$  increases, the distribution of  $\theta_c$  values over examinees converges (with probability one) to the standard normal (or logistic) distribution. Correspondingly the ability  $\theta_{[PN]}$ , which exceeds just a given proportion  $P$  of the abilities  $\theta_c$  in a sample of  $n$  examinees, converges (with probability one), as  $n$  increases, to  $\theta_P = \Phi^{-1}(P)$  [or to  $\Psi^{-1}(P)/D$ ]. This second limit is the ability that exceeds just the proportion  $P$  of abilities in the population. Let  $P_c$  denote the proportion of raw scores in the sample that are less than  $x_c$ , and let

$$\theta(x_c) = \Psi^{-1}(P_c)/D. \quad (17.9.2)$$

Then it follows, under the conditions on item parameters mentioned in the preceding paragraph, that  $\theta(x_c) \rightarrow \theta_c$  (with probability one) as both  $n$  and  $N$  increase. Thus, in practice, with  $N$  and  $n$  finite, we may regard  $\theta(x_c)$  as an estimate of  $\theta_c$ . In the next paragraphs, we treat the  $\theta_c$  as known, with the understanding that in applications they shall be replaced by their numerical estimates  $\theta(x_c)$ .

The likelihood function  $L$  now has as unknown arguments just the  $2n$  item parameters  $a_g$  and  $b_g$ . The maximum likelihood equations

$$\frac{\partial \log L}{\partial a_g} = 0, \quad \frac{\partial \log L}{\partial b_g} = 0,$$

are easily simplified to

$$\frac{1}{N} \sum_{c=1}^N \theta_c \Psi[D a_g(\theta_c - b_g)] = t_g, \quad g = 1, \dots, n, \quad (17.9.3)$$

$$\frac{1}{N} \sum_{c=1}^N \Psi[D a_g(\theta_c - b_g)] = s_g, \quad g = 1, \dots, n, \quad (17.9.4)$$

where

$$s_g = \frac{1}{N} \sum_{c=1}^N u_{gc} \quad \text{and} \quad t_g = \frac{1}{N} \sum_{c=1}^N \theta_c u_{gc}.$$

For each  $g$ , the pair of equations (17.9.3) and (17.9.4) in  $a_g$  and  $b_g$  can be solved for the maximum likelihood estimates  $\hat{a}_g$  and  $\hat{b}_g$  by numerical iteration with the aid of Berkson's (1957) tables of  $\Psi$ .

After each cycle, or after several cycles, of calculation of the successive approximation values

$$[a_g^{(1)}, b_g^{(1)}], \dots, [a_g^{(r)}, b_g^{(r)}], \quad g = 1, \dots, n,$$

the first trial values

$$\theta_c^{(1)} = \theta(x_c) \quad (17.9.5)$$

given by (17.9.2) may be replaced by the successive approximations  $\hat{\theta}_c^{(r)}$ , for  $c = 1, \dots, n$ , where  $\hat{\theta}_c^{(r)}$  is a formal solution of the equation for estimation of  $\theta_c$  when all item parameters are assumed known. This formal solution and its conditions are discussed in detail in Section 20.3 below and used in the next paragraph.

2. Dropping now the assumption made in (1) of a known prior distribution of abilities, we may obtain from  $L$  the maximum likelihood estimates  $\hat{\theta}_c$  of the examinees' abilities  $\theta_c$ , along with the estimates  $\hat{a}_g$  and  $\hat{b}_g$  of item parameters. Even in this case it is convenient to begin an iterative procedure for computing all  $\hat{\theta}_c$ ,  $\hat{a}_g$ , and  $\hat{b}_g$  with first-cycle values  $\theta_c^{(1)} = \theta(x_c)$  defined as in (17.9.2). Then second-cycle values  $\theta_c^{(2)}$  can be obtained from the maximum likelihood equation (see Section 20.3)

$$\partial \log L / \partial \theta_c = 0,$$

or

$$\sum_{g=1}^n a_g \Psi [D a_g (\theta_c - b_g)] = \sum_{g=1}^n a_g u_{gc}, \quad (17.9.6)$$

with  $a_g$  and  $b_g$  replaced by  $a_g^{(1)}$  and  $b_g^{(1)}$ . Then  $\hat{\theta}^{(1)}$  can be replaced by  $\hat{\theta}^{(2)}$  in (17.9.3) and (17.9.4), and the second-cycle values  $a_g^{(2)}$  and  $b_g^{(2)}$  can be obtained as solutions of those equations. Further cycles could run through (17.9.6), (17.9.3), and (17.9.4) in several possible patterns of iteration.

Lord (1967) has successfully applied a procedure similar to that just outlined to various sets of data, using a computer program written by Diana Lees. In one application, the  $a_g$ ,  $b_g$ , and  $\theta_c$  values were simultaneously estimated for 3000 examinees and 90 items (a total of 270,000 item responses). Bock (1967) has reported successful estimation of  $a_g$  and  $b_g$  values by a method based on the assumption that  $\theta_c$  is normally distributed in the population of examinees. Substantial variation in  $a_g$  values was found in both of these applications.

## 17.10 Validity of Test Models

Some aspects of questions of validity and adequacy of fit of specific test models were discussed in Chapter 16. For the logistic model, the estimation methods indicated above may be useful as part of an empirical test of fit. Where specific

techniques of testing fit are concerned, the reader should be aware that some established approaches to testing goodness of fit have come to be considered unsound and potentially misleading by a number of statisticians and scientific workers. An alternative perspective on testing adequacy of models is one based primarily on rather direct, often graphical, comparisons of data with significant aspects of models. Here a crucial role is played by relatively unformalized judgments that involve both the subject-matter context and statistical considerations. Bush (1963) has described and illustrated one such perspective on testing models.

The bearing of some of these questions on statistical efficiency of estimation of ability will be discussed in Section 19.1.

### References and Selected Readings

- BERKSON, J., A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, 1953, **48**, 565-599.
- BERKSON, J., Tables for the maximum likelihood estimate of the logistic function. *Biometrics*, 1957, **13**, 28-34.
- BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58-16*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January 1957.
- BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)
- BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)
- BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics*, 1958, **29**, 1284 (abstract). (c)
- BIRNBAUM, A., Statistical theory of tests of a mental ability. *Op. cit.*, 1285 (abstract). (d)
- BIRNBAUM, A., A unified theory of estimation, I. *Annals of Mathematical Statistics*, 1961, **32**, 112-135. (a)
- BIRNBAUM, A., The theory of statistical inference. New York: Institute of Mathematical Sciences, New York University, 1961. (b) (Mimeographed)
- BIRNBAUM, A., Statistical theory for logistic mental test models with a prior distribution of ability. *Research Bulletin 67-12*. Princeton, N.J.: Educational Testing Service, 1967.
- BOCK, R. D., Fitting a response model for  $n$  dichotomous items. Paper read at the Psychometric Society Meeting, Madison, Wisconsin, March 1967.
- BUSH, R. B., *Handbook of mathematical psychology*, Vol. 1, Chapter 8: Estimation and evaluation. New York: Wiley, 1963.

- COCHRAN, W. G., Estimation of bacterial densities by means of the most probable number. *Biometrics*, 1950, **6**, 105–116.
- CRAMÉR, H., *Mathematical methods of statistics*. Princeton, N.J.: Princeton University Press, 1946.
- FINNEY, D. J., The application of probit analysis to the results of mental tests. *Psychometrika*, 1944, **9**, 31–39.
- FINNEY, D. J., *Probit analysis*. London: Cambridge University Press, 1952.
- FISHER, R. A., On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* (A), 1922, **222**, 309–368. (Reprinted in R. A. Fisher, *Contributions to mathematical statistics*. New York: Wiley, 1950.)
- GUTTMAN, L., Chapters 2, 3, 6, 8, 9 in S. A. Stouffer *et al.*, *Measurement and prediction*. Princeton, N.J.: Princeton University Press, 1950.
- HALEY, D. C., Estimation of the dosage mortality relationship when the dose is subject to error. *Technical Report No. 15*, August 29, 1952. Stanford, Calif.: Contract No. ONR-25140, Applied Mathematics and Statistics Laboratory, Stanford University.
- KEMPTHORNE, O., *An introduction to genetic statistics*. New York: Wiley, 1957.
- LAZARSFELD, P., Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science*, Vol. 3. New York: McGraw-Hill, 1959.
- LINDGREN, B. W., *Statistical theory*. New York: Macmillan, 1960, 1962.
- LOÈVE, M., *Probability theory*. New York: Van Nostrand, 1955.
- LORD, F. M., A theory of test scores. *Psychometric Monograph*, No. 7. Chicago: University of Chicago Press, 1952. (a)
- LORD, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17**, 181–194. (b)
- LORD, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18**, 57–76.
- LORD, F. M., An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Research Bulletin 67-34*. Princeton, N.J.: Educational Testing Service, 1967.
- MCNEMAR, Q., *Psychological statistics*. New York: Wiley, 1962.
- RAO, C. R., *Linear statistical inference and its applications*. New York: Wiley, 1965.
- RASCH, G., *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielson and Lydiche (for Danmarks Paedagogiske Institut), 1960.
- STEVENS, W. L., Control by gauging. *Journal of the Royal Statistical Society* (B), 1948, **10**, 54–108.
- TORGERSON, W. S., *Theory and methods of scaling*. New York: Wiley, 1958.
- TUCKER, L. R., Maximum validity of a test with equivalent items. *Psychometrika*, 1946, **11**, 1–14.
- TUCKER, L. R., Academic ability test. *Research Memorandum 51-17*. Princeton, N.J.: Educational Testing Service, 1951.

## CHAPTER 18

# TEST SCORES, SUFFICIENT STATISTICS, AND THE INFORMATION STRUCTURES OF TESTS

### 18.1 Sufficient Statistics : Definition and Interpretation

In applications, the test models discussed in the preceding chapter principally serve as frames of reference for analyzing and interpreting the information that an examinee's response pattern  $\mathbf{v}$  provides about his ability level  $\theta$ . In this chapter and in Chapters 19 and 20, we shall discuss some general concepts that give exact meaning to several common-sense notions of information. These concepts in turn will guide our consideration of specific applied techniques.

We introduce these concepts on the assumption that a given test model can be known to be valid. In practice we never know whether a given model is precisely valid, and it seems doubtful on general grounds that any model that has one of the special forms considered in the preceding chapter would be precisely valid in any application. Hence it is important to consider the typical conditions under which these concepts and their implications are applied. Under such conditions there is typically only incomplete statistical and theoretical evidence, and this evidence can therefore support only the approximate validity of models. Such considerations have not yet been developed systematically in test theory. Indeed, even for the standard statistical problems of combination and adjustment of observations, some of these considerations have been taken up systematically only in the present decade; notably, development of *efficiency-robust* methods of estimation. Thus our presentation will largely be concerned with what may be called classical statistical concepts and techniques. We may expect these to have some permanent and general value for theory and practice, notwithstanding the indicated need for further basic developments.

The greatest possible simplification of data without loss of information is an important goal in many areas of applied statistics, and particularly in the theory and practice of test scoring. A given set of statistical observations, such as response patterns, may be considered simplified if it is represented by suitable statistics that are easier to use and interpret than the original data. Technically a *statistic* is defined as any function of a response pattern, possibly a vector-valued function. Any score formula  $t = t(\mathbf{v})$  is an example of a (real-valued) statistic.

Any score formula  $t(\mathbf{v})$  seems to provide real simplification, since its range is one-dimensional while the range of  $\mathbf{v}$  is  $n$ -dimensional. However, one hopes that we may determine score formulas that preserve all the information in response patterns, and also order the response patterns appropriately, that is, according to apparently increasing ability. We shall consider the latter desideratum in detail in Chapter 19. In this chapter we consider the former, simplification, in a general and basic way, in terms of statistics that are not necessarily score formulas or real-valued statistics.

The concept of a sufficient statistic, as we shall develop it here, is a precise version of familiar notions of the information in a message or in a set of data. In a given context, it may be that many detailed aspects of a message or a set of data are irrelevant: If they were changed or ignored, there would be no change or loss of information. The concept of a *sufficient statistic* formalizes the notion of abbreviations and deletions that entail no loss of information. The concept of a *minimal sufficient statistic* formalizes the notion of the greatest simplification possible without loss of information.

We shall see that the simple score formula  $\bar{y}$  (proportion correct) is a minimal sufficient statistic in exactly two cases: (1) test models that have equivalent items and a common ICC of any form, and (2) the general logistic test model.

Taking any statistic  $\mathbf{s} = \mathbf{s}(\mathbf{v})$ , we can rewrite the mathematical model of the test:

$$\text{Prob} (\mathbf{V} = \mathbf{v} | \theta) = \text{Prob} [\mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}) | \theta] \text{Prob} [\mathbf{V} = \mathbf{v} | \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}), \theta]. \quad (18.1.1)$$

This is a special case of the general probability formula

$$\text{Prob} (A \text{ and } B | \theta) = \text{Prob} (A | \theta) \text{Prob} (B | A, \theta),$$

which is obtained by considering  $A$  to be the event  $\mathbf{s}(\mathbf{V}) = \mathbf{s}(\mathbf{v})$  and by considering  $B$  to be the event  $\mathbf{V} = \mathbf{v}$ , and by noting that in this case both  $A$  and  $B$  are true if and only if  $B$  is true, and therefore that

$$\text{Prob} (A \text{ and } B | \theta) = \text{Prob} (B | \theta).$$

For some, but not all, statistics  $\mathbf{s}(\mathbf{v})$ , the conditional probability  $\text{Prob} [\mathbf{V} = \mathbf{v} | \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}), \theta]$  is found to be independent of  $\theta$  for each possible  $\mathbf{v}$ . With such a statistic, we can denote the conditional pdf simply by

$$\text{Prob} [\mathbf{V} = \mathbf{v} | \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v})].$$

Equation (18.1.1) then assumes the form

$$\text{Prob} (\mathbf{V} = \mathbf{v} | \theta) = \text{Prob} [\mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}) | \theta] \text{Prob} [\mathbf{V} = \mathbf{v} | \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v})]. \quad (18.1.2)$$

Any statistic  $\mathbf{S}(\mathbf{V})$  for which this holds is called a *sufficient statistic*.

To illustrate and justify the interpretation and application of sufficient statistics, we may think of the observation of a subject's response pattern  $\mathbf{V} = \mathbf{v}$  as being carried out in two stages. First an observation  $\mathbf{s}$  is taken of just the value of a given statistic:  $\mathbf{S} = \mathbf{S}(\mathbf{V})$ . Then an observation is taken from among those response patterns  $\mathbf{v}$  compatible with the observed value  $\mathbf{s}(\mathbf{v}) = \mathbf{s}$ . Mathematical models of these two hypothetical stages in the observation of  $\mathbf{V} = \mathbf{v}$  are in fact given by the respective factors of  $\text{Prob}(\mathbf{V} = \mathbf{v} | \theta)$  in its rewritten form above (Eq. 18.1.2). We may interpret the mathematical condition of sufficiency of  $\mathbf{S}(\mathbf{V})$ , that the second factor be independent of  $\theta$ , to mean that just the value  $\mathbf{s}(\mathbf{v}) = \mathbf{s}$  (but not  $\mathbf{v}$  itself) is reported as it is determined from the response pattern  $\mathbf{v}$  of a subject; for example,

$$\mathbf{v}' = (\mathbf{u}_1, \mathbf{u}_2) \quad \text{and} \quad \mathbf{s}(\mathbf{v}') = \mathbf{u}_1 + \mathbf{u}_2 = \mathbf{1}.$$

As an alternative to the further precise determination of an examinee's response pattern, consider spinning a roulette wheel with outcome labels  $\mathbf{v}$  having respective known probabilities  $\text{Prob}[\mathbf{V} = \mathbf{v} | \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v})]$  independent of  $\theta$ . For example,

$$\text{Prob}[\mathbf{V}' = (1, 0) | U_1 + U_2 = 1] = \frac{1}{3};$$

hence

$$\text{Prob}[\mathbf{V}' = (0, 1) | U_1 + U_2 = 1] = \frac{2}{3}.$$

The roulette wheel and its observed outcome  $\mathbf{v}'$  seem obviously irrelevant to inferences about  $\theta$ , since  $\theta$  does not now play a part in determining  $\mathbf{V}' = \mathbf{v}'$ . However, this merely illustrates vividly that the further determination of the subject's response pattern  $\mathbf{v}'$ , after determination of the value of  $\mathbf{s}(\mathbf{v}')$ , is similarly irrelevant. Thus, when  $\mathbf{S}(\mathbf{V})$  is a sufficient statistic, the absence of mathematical dependence of

$$\text{Prob}[\mathbf{V}' = \mathbf{v}' | \mathbf{S}(\mathbf{V}') = \mathbf{s}(\mathbf{v}')] =$$

on  $\theta$  characterizes the *irrelevance*, for statistical inferences concerning  $\theta$ , of an observation on  $\mathbf{V}'$ , given the observed value of  $\mathbf{s}(\mathbf{V}')$ . We may say that for given  $\mathbf{S}(\mathbf{V}')$ ,  $\mathbf{V}'$  is not an indicant of  $\theta$ .

This concept might be illustrated in further concrete detail by showing that any estimator or classification rule based on  $\mathbf{v}'$  can be matched exactly in all its error-probability properties (in fact, in all probability properties) by an estimator or classification rule based on the statistic  $\mathbf{S} = \mathbf{S}(\mathbf{V})$  and not otherwise dependent on  $\mathbf{V}$ .

A statistic  $\mathbf{s}(\mathbf{v})$  is called *minimal sufficient* if it is sufficient, and if it is a single-valued function

$$\mathbf{S}(\mathbf{V}) = \mathbf{t}[\mathbf{Z}(\mathbf{V})]$$

of every other sufficient statistic. Thus a minimal-sufficient statistic assumes common values on the largest possible sets of points  $\mathbf{v}$  compatible with sufficiency. Such a statistic always exists and is unique, except for one-to-one transformations. (The models considered here, which have discrete sample

spaces, always satisfy the assumptions that guarantee existence; the reader should see Lindgren, 1962, pp. 196–200, for example.) Hence we may refer to *the minimal sufficient statistic*.

## 18.2 Conditions for Sufficiency of a Statistic

We observe that if  $S(V)$  is any sufficient statistic, then the factored form of the probability function of  $V$ ,

$$\text{Prob } (V = v | \theta) = \text{Prob } [S(V) = s(v) | \theta] \text{Prob } [V = v | s(V) = s(v)],$$

contains only one factor that is dependent on  $\theta$ , and this factor is not dependent on  $v$  except through the value of  $s(v)$ . This observation provides us with a convenient criterion for the sufficiency of a statistic. Conversely, let us suppose that  $s(v)$  is any statistic satisfying the condition that the probability function of  $V$  can be written in a factored form in which the factor dependent on  $\theta$  is not dependent on  $v$  except through the value of  $s(v)$ :

$$\text{Prob } (V = v | \theta) = g[s(v), \theta]h(v). \quad (18.2.1)$$

Then from (18.1.1) we have

$$\begin{aligned} \text{Prob } [V = v | S(V) = s(v), \theta] &= \frac{\text{Prob } (V = v | \theta)}{\text{Prob } [S(V) = s(v) | \theta]} \\ &= \frac{\text{Prob } (V = v | \theta)}{\sum_{s(v^*)=s(v)} \text{Prob } (V = v^* | \theta)} \\ &= \frac{g[s(v), \theta]h(v)}{\sum_{s(v^*)=s(v)} g[s(v), \theta]h(v^*)} \\ &= \frac{h(v)}{\sum_{s(v^*)=s(v)} h(v^*)}, \end{aligned}$$

which is independent of  $\theta$ , and we see that  $S(V)$  is therefore sufficient. *Thus (18.2.1) is a convenient necessary and sufficient condition for the sufficiency of a statistic.*

The following simple lemmas will be of use.

**Lemma 18.2.1.** A given statistic is sufficient when the parameter is restricted to any range consisting of an *arbitrary* pair of points if and only if it is sufficient when the parameter is unrestricted.

*Proof.* We note that the condition “no variation of a conditional probability as  $\theta$  varies unrestrictedly” is equivalent to the condition “no variation of a conditional probability as  $\theta$  varies over the values  $\theta_1, \theta_2$ ” considered respectively for all pairs  $\theta_1, \theta_2$  in the range of  $\theta$ .  $\square$

**Lemma 18.2.2.** If a given statistic  $S$  is sufficient, and is minimal sufficient when the range of  $\theta$  is restricted to two specified points  $\theta_1, \theta_2$ , then it is minimal sufficient.

*Proof.* Note that if  $s$  were not minimal sufficient, it would not be possible to write it as a function of some statistic  $z$ , which is sufficient for  $\theta$  unrestricted. However, by Lemma 18.2.1,  $z$  is also sufficient for  $\theta = \theta_1, \theta_2$ , contradicting the minimal sufficiency of  $s$  under this restriction.  $\square$

### 18.3 Test Scores and Sufficient Statistics

We shall give examples of sufficient statistics based on response patterns for the following test models:

1. *Tests with equivalent items.* The items of such tests have a common item characteristic curve  $P_1(\theta)$  of any form

$$\begin{aligned} \text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta) &= \prod_{g=1}^n P_1(\theta)^{u_g} Q_1(\theta)^{1-u_g} \\ &= P_1(\theta)^{\sum_{g=1}^n u_g} Q_1(\theta)^{n - \sum_{g=1}^n u_g} = P_1(\theta)^x Q_1(\theta)^{n-x} \\ &= [P_1(\theta)/Q_1(\theta)]^x Q_1(\theta)^n, \end{aligned} \quad (18.3.1)$$

where  $x = \sum_{g=1}^n u_g$ , denoting the number correct, is the sufficient statistic. This is an example of the general factored form (18.2.1), with the factor  $h(\mathbf{v})$  taking the trivial but admissible form  $h(\mathbf{v}) \equiv 1$ .

2. *Logistic test model.* From (17.2.7), we have

$$\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta) = \left[ \prod_{g=1}^n Q_g(\theta) \right] \exp \left( \theta D \sum_{g=1}^n a_g u_g \right) \exp \left( -D \sum_{g=1}^n a_g b_g u_g \right). \quad (18.3.2)$$

In this factored form we can recognize the weighted-sum score formula

$$x = \sum_{g=1}^n a_g u_g \quad (18.3.3)$$

as a sufficient statistic.

3. *Two-point discrimination problems with any test model.* It is interesting and useful to discuss an oversimplified version of the problem of classification into two ability levels in terms of a test model that may have the general form

$$\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta) = \prod_{g=1}^n P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}, \quad (18.3.4)$$

but whose parameter space is (unrealistically) restricted to consist of just two points,  $\theta = \theta_1$  or  $\theta_2$  ( $\theta_2 > \theta_1$ ). We can then denote the model by

$\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_\alpha)$ , with  $\alpha = 1$  or  $2$  now playing the role of the parameter. Thus

$$\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_\alpha) = \begin{cases} \text{Prob}(\mathbf{V} = \mathbf{v} | \theta_1) & \text{if } \alpha = 1, \\ \text{Prob}(\mathbf{V} = \mathbf{v} | \theta_2) & \text{if } \alpha = 2, \end{cases}$$

or

$$\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_\alpha) = \text{Prob}(\mathbf{V} = \mathbf{v} | \theta_1) \left[ \frac{\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_2)}{\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_1)} \right]^{\alpha-1}, \quad \alpha = 1 \text{ or } 2. \quad (18.3.5)$$

Letting  $L(\mathbf{v}) = \text{Prob}(\mathbf{V} = \mathbf{v} | \theta_2)/\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_1)$ , we have

$$\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_\alpha) = \text{Prob}(\mathbf{V} = \mathbf{v} | \theta_1)L(\mathbf{v})^{\alpha-1}, \quad \alpha = 1 \text{ or } 2. \quad (18.3.6)$$

Here the factor depending on the parameter  $\alpha$  is independent of  $\mathbf{v}$  except through  $L(\mathbf{v})$ , the *likelihood ratio statistic*, which is therefore a sufficient statistic. We have

$$\begin{aligned} \log L(\mathbf{v}) &= \log \left[ \frac{\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_2)}{\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_1)} \right] \\ &= \log \prod_{g=1}^n \left[ \frac{Q_g(\theta_2)}{Q_g(\theta_1)} \right] + \log \prod_{g=1}^n \left[ \frac{P_g(\theta_2)Q_g(\theta_1)}{P_g(\theta_1)Q_g(\theta_2)} \right]^{u_g} \\ &= K + \sum_{g=1}^n w_g u_g, \end{aligned} \quad (18.3.7)$$

where  $K$  is independent of  $\mathbf{v}$  and

$$w_g = w_g(\theta_1, \theta_2) = \log \left[ \frac{P_g(\theta_2)Q_g(\theta_1)}{P_g(\theta_1)Q_g(\theta_2)} \right], \quad g = 1, \dots, k. \quad (18.3.8)$$

Thus the weighted-sum test score formula

$$x(\mathbf{v}) = \sum_{g=1}^n w_g u_g,$$

with *best weights*  $w_g$  defined by (18.3.8), is also a sufficient statistic.

We can readily verify that  $L(\mathbf{v})$  is in fact the minimal sufficient statistic as defined at the end of Section 18.1. It suffices to use (18.3.6) to write the usual formula for the probability of one value of  $L(\mathbf{v})$ , conditionally on the occurrence of either this or one other value; and to observe that this conditional probability depends on that parameter. It follows that no statistic taking the same value over the two sets of points  $\mathbf{v}$  on which  $L(\mathbf{v})$  takes distinct values can be sufficient.

As we noted at the end of Section 18.1, if  $L(\mathbf{v})$  is minimal sufficient, then so is any one-to-one transformation of it, such as  $\log L(\mathbf{v}) - K = \sum_{g=1}^n w_g u_g$ .

In the logistic case, we find that  $w_g(\theta_1, \theta_2) = Da_g(\theta_2 - \theta_1)$ , or

$$\sum_{g=1}^n w_g u_g = D(\theta_2 - \theta_1) \sum_{g=1}^n a_g u_g.$$

When  $D$ ,  $\theta_1$ , and  $\theta_2$  are fixed, this is a one-to-one transformation of  $x = \sum_{g=1}^n a_g u_g$ , the sufficient statistic considered in (2) above. This means that  $x$  is a minimal-sufficient statistic when  $\theta$  is restricted to the values  $\theta_1$  and  $\theta_2$ , and is sufficient for unrestricted  $\theta$ . Since these are the conditions for Lemma 18.2.2, we may conclude that

$$x = \sum_{g=1}^n a_g u_g$$

is *minimal sufficient for the logistic model* when the range of  $\theta$  is unrestricted.

#### 18.4 Sufficiency and the Logistic Test Model

In this section, we shall demonstrate a kind of converse to the result just shown: We shall show that *only* the class of logistic test models admits weighted-sum statistics that are minimal sufficient whether the range of  $\theta$  is restricted or unrestricted.

Finding a weighted-sum statistic that is merely sufficient is not significant. To illustrate this, consider, for example, the rather odd but possible weights

$$w_1 = 0.1, \quad w_2 = 0.01, \dots, w_g = 10^{-g}, \dots.$$

These determine a single weighted-sum scoring formula that is sufficient for *every* test model! To verify this, we need only observe that we can uniquely determine the response pattern  $\mathbf{v}$  from any given value of the statistic

$$x(\mathbf{v}) \equiv \sum_{g=1}^n w_g u_g = \sum_{g=1}^n 10^{-g} u_g.$$

If, for example,  $x(u_1, \dots, u_n) = 0.101$ , we can *deduce* that  $u_1 = 1$ ,  $u_2 = 0$ , and  $u_3 = 1$ ; and if  $n$  exceeds 3, then  $u_g = 0$  for all  $g > 3$ . In fact these weights are devised so that  $x$  is just the decimal fraction with respective digits  $u_1, u_2, \dots$ .

The characterization of the general logistic model given by the theorem and corollaries below is related to conditions, on more general classes of models, for existence of sufficient statistics having certain simple properties (see the discussions of Lindgren, 1962, p. 201, and the general exponential class in Lehmann, 1959, p. 51).

We must appreciate the role that the possible transformations of the ability scale discussed in Section 17.6 may play in any interpretation of this characterization of the logistic model. For example, each test with equivalent items can be given the logistic form by a rescaling of ability: In the example of  $P_1(\theta)$

having the three-parameter logistic form (which is *not* a logistic form), we have

$$P_1(-\infty) = C_1 > 0,$$

and

$$P_1(\theta) > C_1 \quad \text{for all } \theta > -\infty.$$

Solving the relation  $\Psi(\theta^*) = P_1(\theta)$  for  $\theta^*$ , we obtain

$$\theta^* = \theta^*(\theta) = \log [P_1(\theta)/Q_1(\theta)]$$

as a rescaling transformation that gives  $P_1^*(\theta^*) = \Psi(\theta^*)$ . The new  $\theta^*$  ability scale necessarily has a lower bound  $\theta^*$ , determined by  $P_1^*(\theta^*) = \Psi(\theta^*) = C_1$ ; this is *entailed* by the original assumption of the ICC form, in which  $P_1(\theta) > C_1$  for all  $\theta > -\infty$ .

The condition that a model has a logistic form, allowing for possible rescaling of ability, may be expressed thus:

There exists an ability scale  $\theta^* = \theta^*(\theta)$  on which the ICCs  $P_g(\theta)$  of items assume the logistic form

$$P_g^*(\theta^*) \equiv \Psi(A_g\theta^* + B_g) = P_g(\theta), \quad g = 1, \dots, k,$$

where  $A_g$  and  $B_g$  are constants.

The scale transformation required may be written

$$\theta^*(\theta) = A \log \frac{P_h(\theta)}{Q_h(\theta)} + B,$$

where  $P_h(\theta)$  is the ICC of any selected item, and  $A = 1/A_h$  and  $B = -B_h/A_h$  are any constants,  $A > 0$ . We conclude that a model is logistic, up to possible rescaling of  $\theta$ , if and only if there exists a transformation  $\theta^*(\theta)$  such that for each item  $g$ ,  $\log [P_g(\theta)/Q_g(\theta)]$  is a linear function of  $\theta^*$ .

**Theorem 18.4.1.** For  $g = 1, 2$ , let the  $P_g(\theta)$  be any continuous strictly increasing functions defined for  $\theta' < \theta < \theta''$ , with  $0 < P_g(\theta) < 1$ . Let  $\theta_1$  be any fixed value of  $\theta$  in the interval, and let

$$R(\theta) = w_2(\theta_1, \theta)/w_1(\theta_1, \theta) \quad \text{for } \theta \neq \theta_1,$$

where

$$w_g(\theta_1, \theta) = v_g(\theta) - v_g(\theta_1) \quad \text{and} \quad v_g(\theta) = \log [P_g(\theta)/Q_g(\theta)], \quad g = 1, 2.$$

Then, if  $R(\theta)$  is independent of  $\theta$ , there exists a strictly increasing continuous function  $\theta^* = \theta^*(\theta)$  such that  $P_g^*(\theta^*) \equiv P_g(\theta)$  has the logistic form

$$P_g^*(\theta^*) = \Psi(A_g\theta^* + B_g), \quad g = 1, 2,$$

where  $A_g$  and  $B_g$  are constants.

*Proof.* Let

$$\theta^*(\theta) = w_1(\theta_1, \theta).$$

This is a continuous, strictly increasing function of  $\theta$ . Then we have the required form

$$P_1^*(\theta^*) = \Psi(A_1\theta^* + B_1),$$

with

$$A_1 = 1, \quad B_1 = v_1(\theta_1) \equiv \log [P_1(\theta_1)/Q_1(\theta_1)].$$

Let  $A_2 = R(\theta)$ , by assumption a constant for  $\theta \neq \theta_1$ , and let  $B_2 = v_2(\theta_1)$ . Then

$$R(\theta) = \frac{v_2(\theta) - v_2(\theta_1)}{w_1(\theta_1, \theta)} = \frac{v_2(\theta) - B_2}{\theta^*} \equiv A_2,$$

or

$$A_2\theta^* + B_2 = v_2(\theta) = \log [P_2(\theta)/Q_2(\theta)]$$

for all values of  $\theta$ . Thus  $P_2^*(\theta^*)$  also has the required form

$$P_2^*(\theta^*) = \Psi(A_2\theta^* + B_2).$$

For  $\theta = \theta_1$ , we may write  $A_2 w_1(\theta_1, \theta) = v_2(\theta) - v_2(\theta_1)$  and the result still holds.  $\square$

**Corollary 18.4.2.** Any test model (or any set of items) whose ICCs determine respective weights  $w_g(\theta_1, \theta_2)$  for two-point problems, the ratios of which are constant as  $\theta_1$  and  $\theta_2$  vary ( $\theta_1 \neq \theta_2$ ), is equivalent to a logistic test model (or a set of items with logistic ICCs).

Our final and sharpest result linking the logistic form with sufficiency is given as

**Theorem 18.4.3.** Suppose that two continuous ICCs  $P_1(\theta)$ ,  $P_2(\theta)$ , give best weights for respective two-point problems, such that

$$R(\theta_1) = \frac{w_2(\theta_2, \theta_1)}{w_1(\theta_2, \theta_1)} < \frac{w_2(\theta_2, \theta_3)}{w_1(\theta_2, \theta_3)} = R(\theta_3),$$

where  $\theta_1 < \theta_2 < \theta_3$ . Then there is a test model whose items have ICCs only of these forms, in which

- 1) the minimal sufficient statistic for one two-point problem is not sufficient for a second two-point problem, and
- 2) the minimal sufficient statistic for the second problem is not minimal for the first.

*Proof.* We shall illustrate the method of proof by discussing a simple example of the hypothesis of the theorem. Suppose that

$$w_1(\theta_1, \theta_2) = w_2(\theta_1, \theta_2) = w_1(\theta_2, \theta_3) = 1, \quad w_2(\theta_2, \theta_3) = 2.$$

Consider the test model with just two items, with ICCs  $P_1(\theta)$  and  $P_2(\theta)$ , respectively. Then, first, the minimal sufficient statistic for the two-point problem concerning  $\theta_1$  and  $\theta_2$  is  $s = s(u_1, u_2) = u_1 + u_2$ , for which  $s(1, 0) = s(0, 1) = 1$ . Second, for the problem concerning  $\theta_2$  and  $\theta_3$ , the minimal sufficient statistic is  $t = t(u_1, u_2) = u_1 + 2u_2$ , for which  $t(1, 0) = 1 < t(0, 2) = 2$ . Thus  $s$  is not sufficient for the second problem, and  $t$  is not minimal for the first.  $\square$

The property of minimal sufficiency is preserved under any one-to-one transformation of a statistic. Thus, if a weighted-sum scoring formula is transformed by a nonlinear one-to-one function, it will lose its weighted-sum form but retain any sufficiency or minimal sufficiency properties it may have had. If the transformation is nonmonotone, however, the particular simple ordering of response patterns  $(u_1, \dots, u_n)$  will usually also be altered. Furthermore, if the transformation has a multidimensional range, then no new (mathematically natural) simple ordering will replace the original one. Thus we see that in relation to sufficiency properties as such, considerations of specific forms of statistics, such as weighted-sums, have no essential substance. In our discussion above, it has therefore been considerations merely of simplicity and convenience that have led us to choose the particular form

$$\log L(\mathbf{v}) - K,$$

which is linear in the  $u_\theta$ , as a *representation* of the minimal sufficient statistic for any two-point problem. Thus sufficiency concepts alone do not include concepts of information about abilities which may be related to an ordering of response patterns  $\mathbf{v}$ , such as may be determined by any real-valued scoring formula  $x(\mathbf{v})$ . Such concepts are discussed, beginning with the treatment of classification rules in Chapter 19 below.

### 18.5 Sufficiency and the Information Structures of Tests

In Section 17.7, we interpreted the distributions  $F(x | \theta)$  of test scores as representing the information structure of a test. We can now give additional theoretical and practical substance to this interpretation. If the test in question admits a real-valued sufficient statistic  $x$ , then the distributions  $F(x | \theta)$  represent fully the information structure of a test in precisely the sense of the concept of sufficient statistics considered in the preceding sections.

In any case, the family of distributions  $F(x | \theta)$  represents the information structure of the test in this practical sense: It characterizes the forms and error-probability properties of estimation and classification methods that can be based on the scoring formula  $x$  in the ways illustrated in Section 17.4. In our further development of estimation and classification methods, however, we must pay particular attention to cases in which there exists no real-valued sufficient statistic that has the relatively simple and convenient properties considered above.

**References and Selected Readings**

- BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58-16*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January, 1957.
- BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)
- BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)
- BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics*, 1958, **29**, 1284 (abstract). (c)
- BIRNBAUM, A., Statistical theory of tests of a mental ability. *Annals of Mathematical Statistics*, 1958, **29**, 1285 (abstract). (d)
- LEHMANN, E., *Testing statistical hypotheses*. New York: Wiley, 1959.
- LINDGREN, B. W., *Statistical theory*. New York: Macmillan, 1960, 1962.

# CLASSIFICATION BY ABILITY LEVELS

## 19.1 Classification Rules for Distinguishing Two Levels of Ability

In this chapter, we shall consider prototype problems in which the information about ability levels in response patterns is used to order or classify examinees according to higher or lower ability. We shall develop some concepts and techniques which represent optimal solutions of such problems. These solutions in turn will provide information about the appropriate choice of score formulas for various applications. These solutions will also provide information on the dependence of these optimal solutions and score formulas on details of assumed underlying test models and details of specified purposes of classification of examinees by ability levels.

We shall consider the *efficiency* of classification rules of the form illustrated in Section 17.4, according to which an examinee's test score is used to determine his classification as *high ability* ( $x > x_0$ ) or *low ability* ( $x \leq x_0$ ). We shall also consider rules of more general form for such applications. "Efficiency" here refers to the error probability functions of such rules, which are illustrated in Fig. 17.4.3.

The natural basic meaning of "efficient" in such a context is "having adequately small probabilities for all relevant kinds of errors". In any situation where adequately small error probabilities of all relevant kinds are provided by an available test and classification rule, further considerations of efficiency may have little value. But in other situations there may be interest in reducing the error probabilities for given tests and in appraising, comparing, selecting, designing, or constructing new tests and classification rules. The more refined considerations of efficiency relevant here may be termed *statistical efficiency*. They arise when we attempt to use a test model to determine classification rules that are optimal in the sense that they minimize relevant error probabilities. They also arise in broader related contexts, including the design of test models for classification purposes.

The properties of statistical efficiency of classification rules and of estimators, and also the sufficiency property considered in the previous chapter, depend rather sensitively on the detailed form assumed for each item characteristic curve and on the product form by which the local independence assumption

is represented. Since in practice the forms of item characteristic curves are imperfectly known, it is of much practical and theoretical interest to determine to what extent a given inference method, derived under a given assumed form of test model, retains its relevant properties when somewhat different forms hold. It is also of interest to develop methods that are robust for relevant properties, that is, methods that are relatively safe in the sense that they are insensitive to variation in forms of models (within indicated limits). The development of *efficiency-robust* statistical inference methods has been undertaken systematically only in recent years, even for the most standard statistical problems, such as point and interval estimation of means. It has not yet been taken up systematically for models of tests such as those considered here; however, we shall discuss several points relevant to robustness of test models in Sections 19.3 and 20.7.

For any given test model that may be represented by a given probability function  $\text{Prob}(\mathbf{V} = \mathbf{v} | \theta)$ , we can conveniently represent any definite rule for classifying examinees as high or low on the basis of their response patterns  $\mathbf{v}$  as a function  $d = d(\mathbf{v})$ , taking the values one (high) and zero (low). Then, for examinees of any given ability level  $\theta$ , we have

$$\begin{aligned}\text{Prob}(\text{high with rule } d | \theta) &\equiv \text{Prob}[d(\mathbf{V}) = 1 | \theta] \\ &= \sum_{d(\mathbf{v})=1} \text{Prob}(\mathbf{V} = \mathbf{v} | \theta).\end{aligned}$$

The last expression can also be written and interpreted as  $\mathcal{E}[d(\mathbf{V}) | \theta]$ , the expected value of  $d(\mathbf{V})$  at level  $\theta$ . For any rules in which, for some score formula  $x(\mathbf{v})$ ,  $d(\mathbf{v}) = 1$  only when  $x(\mathbf{v}) > x_0$ , we have

$$\mathcal{E}[d(\mathbf{V}) | \theta] = 1 - F(x_0 | \theta),$$

where  $F(x_0 | \theta) = \text{Prob}\{x(\mathbf{V}) \leq x_0\}$ .

## 19.2 Two-Point Classification Problems

Consider any test model and any classification rule  $d(\mathbf{v})$ , and let  $\theta_1$  be any specified ability level definitely considered low in a given context of application (for example, the largest such level). Then

$$\mathcal{E}[d(\mathbf{V}) | \theta_1] = \text{Prob}[d(\mathbf{V}) = 1 | \theta_1]$$

is one of the relevant error probabilities that should have a suitably small value, since it is the probability of a person of a certain low ability  $\theta_1$  being classified as high. Similarly, if  $\theta_2$  is any level considered definitely high (for example, the smallest such level), then

$$1 - \mathcal{E}[d(\mathbf{V}) | \theta_2] = \text{Prob}[d(\mathbf{V}) = 0 | \theta_2]$$

is another such relevant error probability. We can denote these, respectively, by

$$\begin{aligned}\alpha &\equiv \alpha(d | \theta_1) \equiv \mathcal{E}[d(\mathbf{V}) | \theta_1] = \text{Prob } [d(\mathbf{V}) = 1 | \theta_1], \\ \beta &\equiv \beta(d | \theta_2) \equiv 1 - \mathcal{E}[d(\mathbf{V}) | \theta_2] = \text{Prob } [d(\mathbf{V}) = 0 | \theta_2].\end{aligned}\quad (19.2.1)$$

We may achieve a minimal  $\alpha(d | \theta_1)$  for any test by adopting the rule that  $d(\mathbf{V}) = 0$ . In fact this gives  $\alpha = 0$ , since this trivial rule never classifies anyone as high and therefore never misclassifies an examinee with low ability  $\theta_1$ . But usually, of course, this ideal minimization of  $\alpha$  is achieved only along with the unacceptable value  $\beta = \beta(d | \theta_2) = 1$ , since the rule described misclassifies each examinee of high level  $\theta_2$ . Similarly  $\beta = 0$  is attainable but usually only along with  $\alpha = 1$ . Thus the rules ordinarily of interest allow small positive probabilities of errors of each kind; among these, however, rules that minimize these probabilities jointly are to be preferred.

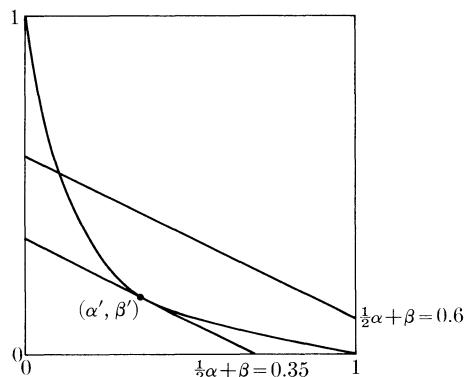


FIG. 19.2.1.  $H(\alpha, \beta) = \frac{1}{2}\alpha + \beta$ . The decision rule that yields values  $\alpha = \alpha'$ ,  $\beta = \beta'$ ,  $\frac{1}{2}\alpha + \beta = 0.35$  is the rule that minimizes  $H(\alpha, \beta)$ . This is demonstrated by the fact that the  $H(\alpha, \beta) = 0.35$  is tangent to the  $(\alpha, \beta)$ -error curve.

We can usefully relate the general goal of determining  $d(\mathbf{v})$  so as to jointly minimize  $\alpha(d | \theta_1)$  and  $\beta(d | \theta_2)$  to the mathematical problem represented in Fig. 19.2.1. Let  $A$  and  $B$  be any given positive numbers, and let

$$H = H(\alpha, \beta) = A\alpha + B\beta = A\alpha(d | \theta_1) + B\beta(d | \theta_2).\quad (19.2.2)$$

This function is strictly increasing in each of the two error probabilities  $\alpha$  and  $\beta$  of interest. Hence the problem of determining the form of  $d(\mathbf{v})$  so as to minimize  $H$  seems at least qualitatively appropriate and relevant to our general goal, and we shall see that it leads to a useful mathematical formulation of our general problem and to a solution for it.\* It is not difficult to show that  $H(\alpha, \beta)$

\* This development is well known in modern elementary mathematical statistics (see, for example, Birnbaum and Maxwell, 1960, pp. 157–159). It includes the Neyman-Pearson lemma (Lindgren, 1962, p. 238) and the derivation of the class of admissible tests between two simple hypotheses (Lindgren, 1962, pp. 162–163).

is minimized by  $d(\mathbf{v})$  if and only if

$$d(\mathbf{v}) = \begin{cases} 1 & \text{if } L(\mathbf{v}) \equiv \frac{\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_2)}{\text{Prob}(\mathbf{V} = \mathbf{v} | \theta_1)} > \frac{A}{B}, \\ 0 & \text{if } L(\mathbf{v}) < \frac{A}{B}. \end{cases} \quad (19.2.3)$$

(See, for example, Birnbaum and Maxwell, 1960, pp. 157–158.) The reader should note that if  $L(\mathbf{v}) = A/B$  for some  $\mathbf{v}$ , then either value can be assigned to  $d(\mathbf{v})$  without affecting the value of  $H(\alpha, \beta)$ . Such rules depend only on the statistic  $L(\mathbf{v})$ , the likelihood ratio statistic. In Section 18.3, we demonstrated that  $L(\mathbf{v})$  is a sufficient statistic only when the range of  $\theta$  is restricted to the values  $\theta_1, \theta_2$  only.

We might have specified the restriction  $B \equiv 1 - A$ ,  $0 < A < 1$ , without restricting the range of  $A/B$ . Then  $H$  could be formally interpreted as a weighted average of the error probabilities  $\alpha$  and  $\beta$ . In a Bayesian approach to inference,  $A$  and  $1 - A$  could then be formally interpreted as respective prior probabilities of the abilities  $\theta_1$  and  $\theta_2$ ; and the minimization of  $H$ , as minimization of the total probability of error.

We may state a basic property of any rule  $d(\mathbf{v})$  that minimizes such a function  $H$ , defined by any choice of  $A$  and  $B$ : No other rule  $d^* = d^*(\mathbf{v})$  can have a smaller error probability of one kind unless it has a larger error probability of the other kind. This property of such rules is called *admissibility*. The proof of the stated property follows immediately: If  $H$  is minimized by  $d(\mathbf{v})$ , and  $d^*(\mathbf{v})$  gives, say,  $\alpha(d^* | \theta_1) < \alpha(d | \theta_1)$ , then if  $\beta(d^* | \theta_2) \leq \beta(d | \theta_2)$ , we have

$$A\alpha(d^* | \theta_1) + B\beta(d^* | \theta_2) < A\alpha(d | \theta_1) + B\beta(d | \theta_2),$$

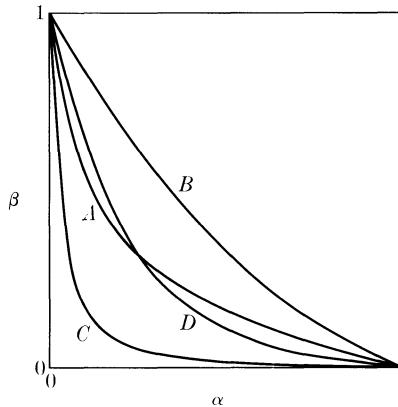
which, by assumption, is a minimum of  $H$ ; but this is a contradiction. We may note that the term “admissible” is also applied to other statistical methods with respect to other efficiency criteria. In this broader context, an admissible method is one whose efficiency cannot be improved for one value of  $\theta$  without some corresponding loss for another value of interest.

In the present context, any admissible rule  $d$  is also a *best* (or *most powerful*) *rule* of level  $\alpha = \alpha(d | \theta_1)$ , in that  $\beta$  is minimized for fixed  $\alpha$ . Since the admissibility property is symmetric in  $\theta_1$  and  $\theta_2$ , it is clear that an analogous term can be applied to  $d$  in relation to  $\beta = \beta(d | \theta_2)$ .

As  $A/B$  increases, fewer points satisfy  $L(\mathbf{v}) \geq A/B$ , and  $\alpha$  decreases while  $\beta$  increases, according to the definition of  $d(\mathbf{v})$  in (19.2.3). In many cases of interest, these successive increments are very small, affording a rather fine choice of  $\alpha$  levels. It can be shown that each  $\alpha$  can be realized exactly by an admissible rule if “randomized” rules are allowed (Lindgren, 1962, pp. 240–241).

Each rule  $d(\mathbf{v})$  may be represented conveniently by a point

$$(\alpha, \beta) = [\alpha(d | \theta_1), \beta(d | \theta_2)]$$

FIG. 19.2.2. Four hypothetical  $(\alpha, \beta)$  curves.

in the unit square, as in the schematic Fig. 19.2.2. For any given test model, the admissible rules are thus represented by respective  $(\alpha, \beta)$ -points, which, it can be shown, constitute a convex curve such as  $A$ , Fig. 19.2.2 (Lindgren, 1962, pp. 232ff). Rules essentially different from the admissible ones have error-probability points falling in the region strictly above curve  $A$ . For example, if the model is such that the best weights  $w_g(\theta_1, \theta_2)$  for the items  $g$  are not all equal, then the classification rules based on the unweighted score  $x(\mathbf{v}) = \sum_{g=1}^n w_g$ , classifying high when its value exceeds a given constant  $x_0$ , would be inadmissible; the weights would be represented, respectively, by the  $(\alpha, \beta)$ -points of a curve such as  $B$ . Rules represented by the  $(\alpha, \beta)$ -points of  $C$ , and those points of  $D$  that fall below  $A$ , do not exist in connection with the test model for which curve  $A$  represents the admissible rules, because their existence would contradict the admissibility of those rules. Only by increasing the number of items in the given test model, or by adopting an otherwise different test model, can rules be attained that are superior to those of curve  $A$ .

Let

$$x = x(\mathbf{v}) = \log L(\mathbf{v}) - K = \sum_{g=1}^n w_g u_g,$$

where

$$w_g = v_g(\theta_2) - v_g(\theta_1) \quad \text{and} \quad v_g(\theta) = \log [P_g(\theta)/Q_g(\theta)],$$

as in Section 18.4. We note that  $x$  is an increasing function of  $L(\mathbf{v})$ , and from this we see that for two-point problems, with any test model, the admissible rules include all those that have the simple form in which high ability is indicated just by  $x(\mathbf{v}) \geq x_0$  for some number  $x_0$ . The other admissible rules differ from this form only in the ways in which they are defined for the score value  $x(\mathbf{v}) = x_0$ . Thus a rule that indicates high ability just when  $x(\mathbf{v}) > x_0$  is also admissible, and so is a rule that differs from this only by having a more

complicated form (possibly depending on  $\mathbf{v}$  or on an auxiliary randomization variable) just when  $x(\mathbf{v}) = x_0$ . If each single value of  $x(\mathbf{v})$ , such as  $x_0$ , has a very small probability under each  $\theta$ , as is true with many models, then such variations in the definition of  $d(\mathbf{v})$  give alternative admissible rules which are nearly identical in their error-probability properties.

The two-point problem is of interest primarily as a prototype and as a technical step toward more realistic analyses in which all abilities definitely considered either high or low are taken into account. Figure 17.4.3 shows the error probabilities of a rule of the form illustrated above, when the full range of  $\theta$  is considered. Rules having the admissibility property in two-point problems often can be proved admissible also where the range of  $\theta$  is unrestricted. In this situation, an admissible rule is one such that no other rule can have a smaller error probability at one value of  $\theta$  in the range  $\theta \leq \theta_1$  or  $\theta \geq \theta_2$  without having a larger error probability at some other value in the same range.

As an illustration, suppose that we have the problem of classification as high ability,  $\theta \geq \theta_2$ , or low ability,  $\theta \leq \theta_1$  ( $\theta_1 < \theta_2$ ). Consider a rule  $d(\mathbf{v})$  that is admissible for the two-point problem,  $\theta = \theta_1$  or  $\theta = \theta_2$ , and that classifies as high just when  $x(\mathbf{v}) \equiv \log L(\mathbf{v}) - K > x_0$ , with positive error probabilities  $\alpha = \alpha(d | \theta_1)$ ,  $\beta = \beta(d | \theta_2)$ , where  $\alpha = 0.03$ ,  $\beta = 0.08$ , as in Section 17.4.3. Suppose further that  $x_0$  is not among the attainable values of  $x(\mathbf{v})$ . Then ambiguities regarding classification when  $x(\mathbf{v}) = x_0$  disappear, and we see that  $d(\mathbf{v})$  is the *unique* rule that minimizes (19.2.2) for the  $A$  and  $B$  corresponding to  $x_0$ . This, in turn, implies that  $d(\mathbf{v})$  is the only rule with error probabilities  $\alpha$  and  $\beta$  at  $\theta_1$  and  $\theta_2$ , respectively. Since  $d(\mathbf{v})$  is admissible for this two-point problem, we may now conclude that it is admissible in the more general case: Any  $d^*(\mathbf{v})$  with a smaller error probability than  $d(\mathbf{v})$  at some  $\theta' (\theta' \leq \theta_1$  or  $\theta' \geq \theta_2)$  must have a larger error probability at either  $\theta_1$  or  $\theta_2$ .

Two points, such as  $[\theta_1, \alpha(d | \theta_1)]$ ,  $[\theta_2, \beta(d | \theta_2)]$ , on the error probability curve of any admissible rule provide a useful, simple partial description of that rule's properties. At the same time, two such points represent bounds on attainable error probabilities in the sense that these values  $\alpha, \beta$  cannot be simultaneously improved on by any other rule based on the same mental test.

The scoring formula considered in the preceding comments was suggested by its admissibility for the two-point problem. Somewhat different choices of values  $\theta_1, \theta_2$ , chosen to mark a *highest definitely low ability* and a *lowest definitely high ability* schematically, would in general determine somewhat different weights  $w_g$ . Of course the admissibility (or best) property and the results of the preceding discussion will hold again for each such rule.

We shall next consider the question: What is the most general test model for which the *same* weighted-sum scoring formula is obtained in *each* such determination of an admissible rule? Another form of this question is: As  $\theta_1, \theta_2$  vary (subject to  $\theta_1 < \theta_2$ ), what is the most general test model (if one exists) that admits a single classification rule that is best for all values  $\theta_2 > \theta_1$ , a so-called *uniformly best classification rule*? The answer is: The logistic test

model. The proof is based on considerations like those of Section 18.4, and on the observation that when  $\theta_1$  is fixed and  $\theta_2$  varies, if one scoring formula is to remain identically equal to

$$\log L(\mathbf{v}) \equiv \log [\text{Prob} (\mathbf{V} = \mathbf{v} | \theta_2) / \text{Prob} (\mathbf{V} = \mathbf{v} | \theta_1)]$$

[apart from a factor  $K(\theta_2)$ , which may vary with  $\theta_2$  but is independent of  $\mathbf{v}$ ], then the ratios of weights  $w_g/w_h$  in the scoring formula must be independent of variation in  $\theta_2$ .

We have seen here that the two plausible concepts, sufficiency and admissibility, tend to agree with, and in a sense to complement, one another by supporting from somewhat different viewpoints the use of certain statistics (score formulas). Each concept seems to derive further plausibility from appearing compatible with, and complementary to, the other plausible concept. This in turn seems to confirm the plausibility and appropriateness of a general approach to statistical inference problems (including classification, estimation, and testing hypotheses) that incorporates these as basic concepts. Indeed the statistical methods thus obtained seem highly appropriate in some contexts, such as some situations requiring acceptance or rejection of applicants. But it would be a misleading oversimplification to suggest that we have adequately illustrated the significance of these concepts for statistical methodology in general. In fact other aspects of sufficiency and of error probabilities, particularly as concepts of statistical evidence in empirical research contexts, are among the basic concepts of statistical inference currently undergoing critical reappraisal; the reader might see, for example, Savage (1962), Birnbaum (1962, 1968a, 1968b), Novick and Hall (1965), Tukey (1962), Mosteller and Wallace (1964), and Hartigan (1965).

### 19.3 Locally Best Weights and Classification Rules

With many models, the detailed determination of best weights, and of related statistical theory and calculations, can be carried through most conveniently and completely in a special limiting case, namely, the case in which the difference  $\theta_2 - \theta_1$  becomes arbitrarily small. Some detailed treatment of this case will provide convenient illustrations here. Even more importantly, it will be convenient to base the principal methods of estimation to be developed in the next chapter on the techniques developed here.

For  $\theta_2$  sufficiently near  $\theta_1$ , the best weights  $w_g = w_g(\theta_1, \theta_2)$  are effectively and conveniently approximated by the *locally best weights*, defined by

$$w_g \equiv w_g(\theta_1) = \lim_{\theta_2 \rightarrow \theta_1} [w_g(\theta_1, \theta_2) / (\theta_2 - \theta_1)]. \quad (19.3.1)$$

(We note that division of each weight  $w_g(\theta_1, \theta_2)$  by the same positive constant  $(\theta_2 - \theta_1)$  does not change the ratios between weights of respective items; just

these ratios characterize best weights.) Since

$$\frac{w_g(\theta_1, \theta_2)}{\theta_2 - \theta_1} = \frac{v_g(\theta_2) - v_g(\theta_1)}{\theta_2 - \theta_1}, \quad (19.3.2)$$

we have

$$\begin{aligned} w_g(\theta_1) &= \lim_{\theta_2 \rightarrow \theta_1} \frac{v_g(\theta_2) - v_g(\theta_1)}{\theta_2 - \theta_1} = \left. \frac{\partial}{\partial \theta} v_g(\theta) \right|_{\theta=\theta_1} \\ &= \left. \frac{\partial}{\partial \theta} \log \frac{P_g(\theta)}{Q_g(\theta)} \right|_{\theta=\theta_1} = \frac{P'_g(\theta_1)}{P_g(\theta_1)Q_g(\theta_1)}, \end{aligned} \quad (19.3.3)$$

where  $P'_g(\theta) = (\partial/\partial\theta)P_g(\theta)$ , assuming that the derivative exists (as it does in most of our examples). To see that ratios between these weights for items  $g$  and  $h$  are close approximations to ratios between best weights  $w_g(\theta_1, \theta_2)$  and  $w_h(\theta_1, \theta_2)$  when  $\theta_2$  is sufficiently near  $\theta_1$ , observe that when  $\theta_2 - \theta_1$  is sufficiently small, we have (by the definition of a derivative)

$$w_g(\theta_1, \theta_2) \doteq \left. \frac{\partial}{\partial \theta_2} w_g(\theta_1, \theta_2) \right|_{\theta_2=\theta_1} (\theta_2 - \theta_1) \equiv w_g(\theta_1)(\theta_2 - \theta_1),$$

and hence

$$\frac{w_g(\theta_1, \theta_2)}{w_h(\theta_1, \theta_2)} \doteq \frac{w_g(\theta_1)}{w_h(\theta_1)}$$

if  $w_h(\theta_1) \neq 0$  and  $w_h(\theta_1, \theta_2) \neq 0$ .

It can also be shown, as we might expect, that a classification rule based on a score  $t$  with locally best weights  $w_g(\theta_1)$  has, of all rules with the same value of  $\alpha = \alpha(d \mid \theta_1) = 1 - F(t_0 \mid \theta_1)$ , the largest possible value of the derivative at  $\theta_1$  of the error-probability function

$$\left. \frac{\partial}{\partial \theta} [1 - F(t_0 \mid \theta)] \right|_{\theta=\theta_1} = -F'(t_0 \mid \theta_1),$$

provided the derivatives  $P'_g(\theta)$  of the ICCs exist. Such a rule is called *locally best* (at  $\theta_1$ ); the reader might see, for example, Lehmann (1959, p. 364) and Rao (1965, pp. 382–383). The maximum condition means that among all rules with the given error probability  $\alpha$  at  $\theta_1$ , for each positive  $\epsilon$ , no other rule has error probabilities  $\beta$  at least equally small for all  $\theta_2$  in the interval

$$\theta_1 < \theta_2 < \theta_1 + \epsilon.$$

We indicated near the end of Section 19.2 that if a test model has best weights  $w_g(\theta_1, \theta_2)$ , which are independent of  $\theta_1$  and  $\theta_2$  (at least as regards their mutual ratios), then the model is logistic. Now locally best weights  $w_g(\theta_1)$  are essentially (that is, in their ratios) limits as  $\theta_2 \rightarrow \theta_1$  of weights  $w_g(\theta_1, \theta_2)$ ; and it can be shown that such limits are essentially independent of  $\theta_1$  only if the  $w_g(\theta_1, \theta_2)$  are essentially independent of  $\theta_1$  and  $\theta_2$ . Thus a test model has locally best

weights independent of  $\theta_1$  if and only if it is a logistic model. Examples of locally best weights are given below.

### 1. Logistic model

$$w_g(\theta) = Da_g \frac{\psi[DL_g(\theta)]}{\Psi[DL_g(\theta)]} \{1 - \Psi[DL_g(\theta)]\} = Da_g.$$

The final expression is easily obtained using (19.3.3). On canceling the inessential common factor  $D$  of such weights, we obtain the weights  $a_g$ , which have become familiar in other connections, in particular as best weights  $w_g(\theta_1, \theta_2) \equiv a_g$ , independent of  $\theta_1$  and  $\theta_2$ .

### 2. Normal ogive model

$$w_g(\theta) = a_g \frac{\varphi[L_g(\theta)]}{\Phi[L_g(\theta)]\Phi[-L_g(\theta)]}.$$

The factor multiplying  $a_g$  here is the function  $J(s) = \varphi(s)/\Phi(s)\Phi(-s)$ , where  $s = L_g(\theta) = a_g(\theta - b_g)$  depends on  $\theta - b_g$  as well as on  $a_g$ . The behavior of  $J(s)$  is indicated sufficiently for our purposes by a brief table:

$s$	0	$\pm 1$	$\pm 2$	$\pm 3$
$J(s)$	1.6	1.8	2.4	3.4

Thus, for efficient discrimination between ability levels in the neighborhood of a given level  $\theta_1$ , the weighting  $w_g$  to be given to response  $u_g$  will vary by a factor possibly as large as 2, depending upon the item's difficulty  $b_g$  (through the difference  $\theta_1 - b_g$ ).

One starting point for an investigation of robustness of validity and efficiency properties would be a study of error-probability functions of classification rules that are optimal under certain logistic test models when normal-ogive models with corresponding item-parameters are in fact valid; and conversely.

### 3. Three-parameter logistic model

$$w_g(\theta) = Da_g \Psi[DL_g(\theta) - \log c_g].$$

The moments and distributions of composite scores with such weights were discussed in Section 17.7; and applications of such composite scores are discussed throughout Chapters 17 through 21.

## 19.4 More General Classification Rules, Composite Scores, and Statistical Efficiency in General

The considerations of the preceding sections can be extended to problems of classification into three or more ranges of ability; *low*, *middle*, and *high*. Admissible rules for three-point problems, for example, are obtainable on the

basis of a direct generalization of the derivation of admissible two-point rules given above (see, for example, Birnbaum and Maxwell, 1961). No single best weighted score alone will determine all the admissible rules, except in logistic test models. In other models, at least a pair of such statistics jointly, or else a statistic of another form, must be used as the basis for admissible rules. For example, a score with weights  $w_g(\theta_1, \theta_2)$ , and another with weights  $w_g(\theta_2, \theta_3)$ , are jointly sufficient when  $\theta$  is restricted to three values  $\theta_1, \theta_2, \theta_3$  in any test model with  $0 < P_g(\theta_i) < 1$  for all  $i, g$ .

It can be shown that the most general test model for which a single score formula suffices for such purposes is the logistic model. Rules having the simple form determined by two critical values  $x_1$  and  $x_2$ , with scores  $x(\mathbf{v}) \leq x_1$  classified as low, scores  $x(\mathbf{v}) \leq x_2$  classified as high, and others classified as middle, are admissible in each case of the logistic model [when  $x(\mathbf{v}) = \sum_{g=1}^n a_g u_g$ ]. But for essentially different models (with nonequivalent items), rules of this form, based on any score formula  $x(\mathbf{v})$ , are in general inadmissible. The method of proof is a direct extension of that for the two-point problem, which we indicated near the end of Section 19.2.

Another general mode of formulating inference or classification problems, partly related to formulations in terms of error probabilities, is that of statistical decision theory. After specifying the costs or disutilities of errors of each of the numerous kinds possible in a given problem, one can characterize admissible classification rules in such general problems. In problems involving a single real-valued parameter and ordered alternative decisions, it can be shown under certain broad conditions that only in the general logistic test model are all the admissible decision rules based on a single real-valued statistic whose increasing values indicate the respective ordered decisions (see Lindgren, 1962, pp. 205–209, and references therein).

A mathematical condition that characterizes and entails these simple, intuitively natural forms for admissible rules for a broad class of inference and decision problems is the *monotone likelihood ratio property*: For any fixed  $x_1, x_2$  where  $x_2 > x_1$ ,

$$\text{Prob}[x(\mathbf{V}) = x_2 | \theta] / \text{Prob}[x(\mathbf{V}) = x_1 | \theta]$$

is strictly increasing in  $\theta$ . The logistic model and score formula meet this condition. Another closely related condition that holds with the logistic model is that  $\text{Prob}(\mathbf{V} = \mathbf{v} | \theta)$  has the *exponential* (or Koopman-Darmois) form. The latter condition entails both the existence of a sufficient statistic of the weighted sum form and the monotone likelihood ratio property for the distributions of that statistic (see, for example, Lindgren, 1962, pp. 200–202, and references therein).

Rules of these intuitively natural forms are not in general admissible with models essentially different from the logistic. But most thinking about test scores, most practice with them, seems to incorporate, at least tacitly, the assumption that a single suitably defined composite score (usually the un-

weighted sum, occasionally a weighted sum) serves efficiently to indicate respective ordered decisions or inferences in a monotone way. The properties of sufficiency, monotonicity, and statistical efficiency discussed above may be regarded as supplying some *explications* of this tacit assumption. But on these terms, the assumption is given *justification* only under the very restrictive condition that a test model have the logistic form; with models of essentially different form, some loss of efficiency is entailed. We shall give detailed consideration to some quantitative aspects of such losses; for example, in Section 20.8 we shall appraise losses of precision in estimation based on nonoptimal score formulas.

The tacit assumption just mentioned is doubtless related to the plausible but overly simple expectation that efficient statistical classification rules should resemble efficient nonstatistical rules; in particular, that some estimate of an unknown value  $\theta$  should visibly play the role in the classification procedure that would be played by  $\theta$  if it were known. This point about test scores and their uses is in some respects akin to certain subtle questions of general methodology which have been of much interest in the development of mathematical statistics. The latter are basic questions in the logic of measurement in the presence of errors of measurement; they involve the intuitive and the operational aspects of statistical methods for a single unknown parameter, and the relations between efficient estimators, sufficient statistics, and such properties as monotone likelihood ratio and the Koopman-Darmois form.

### 19.5 Quantitative Appraisal and Efficient Design of Classification Rules

Let  $\theta_1$  and  $\theta_2$  be ability levels representing, respectively, specified low and high levels of ability as above. Consider any score formula of the weighted sum form  $x(\mathbf{v}) = \sum_{g=1}^n w_g u_g$ , and any classification rule of the form that classifies as high when  $x > x_0$ , and as low when  $x \leq x_0$ , where  $x_0$  is a specified number (*critical score*). The error probability function of such a rule is

$$1 - F(x_0 | \theta) \quad \text{for } \theta \leq \theta_1, \quad \text{and} \quad F(x_0 | \theta) \quad \text{for } \theta \geq \theta_2,$$

as illustrated in Fig. 17.4.3. We shall focus attention particularly on the error probabilities

$$\alpha = \alpha(d | \theta_1) = 1 - F(x_0 | \theta_1) \quad \text{and} \quad \beta = \beta(d | \theta_2) = F(x_0 | \theta_2), \quad (19.5.1)$$

where  $d = d(\mathbf{v})$  refers to a particular rule of the form described. Here we shall ordinarily use the normal approximation presented in Section 17.7:

$$F(x_0 | \theta) \doteq \Phi \left\{ \frac{x_0 - \mathcal{E}(x(\mathbf{V}) | \theta)}{\sigma[x(\mathbf{V}) | \theta]} \right\}. \quad (19.5.2)$$

For notational simplicity in this section, we shall henceforth replace the approximate equality symbol  $\doteq$ , where it would ordinarily occur with this normal

approximation, by the equality symbol  $=$ . The approximate equality symbol will be used for other relations where appropriate, unless the contrary is specifically stated.

On this basis it is convenient to deduce formally a number of results. These results will provide useful general guidance. Also, in specific problems of interest they will provide quantitative solutions which may be treated as first approximations. In applications these solutions can be considered in connection with appropriate bounds on errors of approximation that may be available, or with other checks.

On these terms, we have

$$\alpha = 1 - \Phi \left\{ \frac{x_0 - \mathcal{E}[x(\mathbf{V}) \mid \theta_1]}{\sigma[x(\mathbf{V}) \mid \theta_1]} \right\} \quad \text{and} \quad \beta = \Phi \left\{ \frac{x_0 - \mathcal{E}[x(\mathbf{V}) \mid \theta_2]}{\sigma[x(\mathbf{V}) \mid \theta_2]} \right\}. \quad (19.5.3)$$

Now

$$\mathcal{E}[x(\mathbf{V}) \mid \theta] = \sum_{g=1}^n w_g P_g(\theta) \quad \text{and} \quad \text{Var}[x(\mathbf{V}) \mid \theta] = \sum_{g=1}^n w_g^2 P_g(\theta) Q_g(\theta). \quad (19.5.4)$$

When we substitute the formulas (19.5.4) in formulas (19.5.3), we obtain formulas (19.5.5) and (19.5.6), which give  $\alpha$  and  $\beta$  explicitly in terms of the item parameters and which also give other specific details of the structure of our test model and classification rule:

$$\alpha = 1 - \Phi \left\{ \frac{x_0 - \sum_{g=1}^n w_g P_g(\theta_1)}{\left[ \sum_{g=1}^n w_g^2 P_g(\theta_1) Q_g(\theta_1) \right]^{1/2}} \right\}, \quad (19.5.5)$$

$$\beta = \Phi \left\{ \frac{x_0 - \sum_{g=1}^n w_g P_g(\theta_2)}{\left[ \sum_{g=1}^n w_g^2 P_g(\theta_2) Q_g(\theta_2) \right]^{1/2}} \right\}. \quad (19.5.6)$$

These relations are the basis for a number of deductions relevant to appraisal and design both of classification rules and of tests to be used with classification rules. At the same time, these considerations constitute a convenient technical step in developing estimation methods, and methods of appraisal of information structures of tests and design of tests, as we shall show in the next chapter.

**Critical score.** Suppose we are given any test model, represented by

$$\text{Prob} (\mathbf{V} = \mathbf{v} \mid \theta) = \prod_{g=1}^n P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}, \quad (19.5.7)$$

and any score formula  $x = \sum_{g=1}^n w_g u_g$ . Consider the problem of finding a

critical score  $x_0$  that determines a classification rule  $d = d(\mathbf{v})$  of the form in which  $x(\mathbf{v}) > x_0$  classifies as high, and such that the error probability  $\alpha(d \mid \theta_1)$  has a specified value  $\alpha$  (for example,  $\alpha = 0.06$ ) at a specified low ability  $\theta_1$ . We require then that

$$\alpha = 1 - \Phi \left\{ \frac{x_0 - \mathcal{E}[x(\mathbf{V}) \mid \theta_1]}{\sigma[x(\mathbf{V}) \mid \theta_1]} \right\}. \quad (19.5.8)$$

This determines the required value

$$\begin{aligned} x_0 &= \mathcal{E}[x(\mathbf{V}) \mid \theta_1] + \Phi^{-1}(1 - \alpha)\sigma[x(\mathbf{V}) \mid \theta_1] \\ &= \sum_g^n w_g P_g(\theta_1) + \Phi^{-1}(1 - \alpha) \left[ \sum_{g=1}^n w_g^2 P_g(\theta_1) Q_g(\theta_1) \right]^{1/2}. \end{aligned} \quad (19.5.9)$$

To determine the error probability of this rule at any specified high ability level  $\theta_2$ , we substitute (19.5.9) in (19.5.6) to obtain

$$\begin{aligned} \beta &= \beta(d \mid \theta_2) \\ &= \Phi \left\{ \frac{\mathcal{E}[x(\mathbf{V}) \mid \theta_1] + \Phi^{-1}(1 - \alpha)\sigma[x(\mathbf{V}) \mid \theta_1] - \mathcal{E}[x(\mathbf{V}) \mid \theta_2]}{\sigma[x(\mathbf{V}) \mid \theta_2]} \right\} \\ &= \Phi \left\{ \Phi^{-1}(1 - \alpha) \frac{\sigma[x(\mathbf{V}) \mid \theta_1]}{\sigma[x(\mathbf{V}) \mid \theta_2]} + \frac{\mathcal{E}[x(\mathbf{V}) \mid \theta_1] - \mathcal{E}[x(\mathbf{V}) \mid \theta_2]}{\sigma[x(\mathbf{V}) \mid \theta_2]} \right\}, \end{aligned} \quad (19.5.10)$$

into which we can further substitute the summation formulas (19.5.4) for the moments of  $x(\mathbf{V})$ , when required.

**Required number of items.** In Sections 18.2 through 18.4, we have seen that for any test with equivalent items, the admissible rules are based on equal weighted composite scores  $x(\mathbf{v}) = \sum_{g=1}^n u_g$ . We have

$$\mathcal{E}[x(\mathbf{V}) \mid \theta] = nP(\theta), \quad \sigma^2[x(\mathbf{V}) \mid \theta] = nP(\theta)Q(\theta), \quad (19.5.11)$$

and

$$\beta = \Phi \left\{ \Phi^{-1}(1 - \alpha) \left[ \frac{P(\theta_1)Q(\theta_1)}{P(\theta_2)Q(\theta_2)} \right]^{1/2} + n^{1/2} \frac{P(\theta_1) - P(\theta_2)}{[P(\theta_2)Q(\theta_2)]^{1/2}} \right\}. \quad (19.5.12)$$

If  $P(\theta_1) - P(\theta_2) < 0$  when  $\theta_2 > \theta_1$ , as we have usually assumed (and have illustrated by most of the models of Chapter 17), we see that  $\beta$  decreases to zero as the number  $n$  of items increases, regardless of any more specific features of the common item characteristic curve  $P(\theta)$ .

A problem of test design that can be solved explicitly in this case is that of determining the minimal number  $n$  of items required to meet two given bounds  $\alpha^*$  and  $\beta^*$  on error probabilities  $\alpha(d \mid \theta_1) \leq \alpha^*$  and  $\beta(d \mid \theta_2) \leq \beta^*$ . Let us temporarily ignore the fact that  $n$  takes only integral values. We then see that

the last equation gives

$$n = \left\{ \frac{\Phi^{-1}(1 - \alpha^*)[P(\theta_1)Q(\theta_1)]^{1/2} - \Phi^{-1}(\beta^*)[P(\theta_2)Q(\theta_2)]^{1/2}}{[P(\theta_2) - P(\theta_1)]^2} \right\}^2. \quad (19.5.13)$$

The required number of items is the smallest integer greater than or equal to the preceding quantity.

When  $\beta^* = \alpha^*$  has been specified,  $\Phi^{-1}(\beta^*) = -\Phi^{-1}(1 - \alpha^*)$ . Then, on rearranging (19.5.13), we have

$$\frac{\Phi^{-1}(1 - \alpha^*)}{n^{1/2}} = \frac{P(\theta_2) - P(\theta_1)}{[P(\theta_1)Q(\theta_1)]^{1/2} + [P(\theta_2)Q(\theta_2)]^{1/2}}. \quad (19.5.14)$$

The right-hand member of (19.5.14) depends just on the form of the common ICC and constitutes a natural measure of the amount of information per item provided by items of the given form, a measure of information specifically related to the contributions of items toward minimizing common error probabilities in discriminating between levels  $\theta_1$  and  $\theta_2$ .

**Local case: a measure of information.** Of course, the right-hand member of (19.5.14) depends on  $\theta_1$  and  $\theta_2$ . If we divide the right-hand member by  $(\theta_2 - \theta_1)$ , it becomes a kind of information measure expressed on the scale "per unit separation between ability levels". We may now obtain a particularly convenient approximation formula by proceeding to the limit  $(\theta_2 \rightarrow \theta_1)$ , provided that it exists, as indeed it does in most of the examples of ICCs in Chapter 17. This limit is

$$\frac{1}{2} \frac{P'(\theta_1)}{[P(\theta_1)Q(\theta_1)]^{1/2}}, \quad \text{where } P'(\theta) = \frac{\partial}{\partial \theta} P(\theta). \quad (19.5.15)$$

On squaring this expression and deleting the constant factor, we obtain the formulation

$$I(\theta_1, u_\theta) = \frac{P'(\theta_1)^2}{P(\theta_1)Q(\theta_1)} \quad (19.5.16)$$

as a measure of information per item having ICC of the form  $P(\theta)$ , that can be used to discriminate abilities in a neighborhood of  $\theta_1$ . Using the limit approximation in (19.5.14), we have

$$\begin{aligned} \frac{\Phi^{-1}(1 - \alpha^*)}{n^{1/2}} &= \frac{P(\theta_2) - P(\theta_1)}{[P(\theta_1)Q(\theta_1)]^{1/2} + [P(\theta_2)Q(\theta_2)]^{1/2}} \\ &\doteq 2 \frac{P'(\theta_1)}{[2P(\theta_1)Q(\theta_1)]^{1/2}} (\theta_2 - \theta_1), \end{aligned}$$

or

$$nI(\theta_1, u_\theta) \doteq 4 \left[ \frac{\Phi^{-1}(1 - \alpha^*)}{(\theta_2 - \theta_1)} \right]^2. \quad (19.5.17)$$

This formula makes explicit the role of the item information function in characterizing the contribution, per item having given ICC, toward reduction of probabilities of errors concerning close alternative values  $\theta_1$  and  $\theta_2$ . Thus  $\alpha^*$  is lowered the same amount by doubling  $I(\theta_1, u_g)$  as by doubling the number of items.

In the following chapter, we shall see that the item information function  $I(\theta, u_g)$ , represented by (19.5.16) taken as a function of  $\theta$ , is related to the information functions  $I(\theta, x)$  defined in Section 17.7; and that each of these plays a basic role in the theory and techniques of estimation.

**Best difficulty levels.** Let us consider an example of the uses of the general formulas (19.5.10) and (19.5.12) for  $\beta$  in problems of item selection and the design of tests and classification rules. Suppose that

- 1)  $\theta_1, \alpha = \alpha(d | \theta_1)$ ,  $n$ , and  $\theta_2$  ( $\theta_2 > \theta_1$ ) are given, and the scale for  $\theta$  is fixed;
- 2) we may choose any set of  $n$  items with logistic ICCs.

Our problem is to minimize  $\beta(d | \theta_2)$  by appropriate choice of a logistic test design (model). In other words, we must choose a set of the item parameters  $(a_1, b_1, a_2, b_2, \dots, a_n, b_n)$ , and a classification rule, that is, a critical score  $x_0$ , so as to minimize  $\beta$ . We recall (see Section 17.2) that a single logistic item with a large enough value of  $a_g$  would suffice to give arbitrarily small values to  $\alpha$  and  $\beta$ , provided that  $b_g$  lies between  $\theta_1$  and  $\theta_2$ . Experience with test items shows that there is an effective upper bound on the discriminating power of items ordinarily available. Hence we may add an upper bound on the  $a_g$ , say  $a'$ , to our problem, and further simplify and schematize it for illustrative purposes by requiring that the items be equivalent. (It is plausible, and can be shown formally, that a best choice would in fact be one with each  $a_g$  equal to the given upper bound  $a'$  and with some common value  $b$  chosen for the  $b_g$ .) Thus reduced, our problem is to choose  $b$  and  $x_0$  so as to minimize  $\beta$  as given in (19.5.12).

**Approximate determination of best difficulty level for a special case.** If  $\alpha = \alpha(\theta_1) = \frac{1}{2}$ , then  $\Phi^{-1}(1 - \alpha) = 0$ , and we see from the form of (19.5.12) that  $\beta$  is minimized when

$$[P(\theta_2) - P(\theta_1)]/[P(\theta_2)Q(\theta_2)]^{1/2} \quad (19.5.18)$$

is maximized, or, equivalently, when

$$\frac{P(\theta_2) - P(\theta_1)}{\theta_2 - \theta_1} \frac{1}{[P(\theta_2)Q(\theta_2)]^{1/2}} \quad (19.5.19)$$

is maximized. As we saw above, the limit of this expression, as  $\theta_2 \rightarrow \theta_1$ , is

$$\frac{1}{2}\sqrt{I(\theta_1, u_1)}. \quad (19.5.20)$$

If  $\theta_2$  is sufficiently near  $\theta_1$ , then the value  $b$  that maximizes  $I(\theta_1, u_1)$  is an approximate solution of our problem. [Clearly this value of  $b$  also maximizes the slope of  $F(x_0 | \theta)$  at  $\theta = \theta_1$ .]

In Section 20.4, we shall examine and illustrate the forms of item information curves for specific models. We shall show there that for both the logistic and normal ogive models,  $I(\theta, u_g)$  is symmetric and unimodal in  $\theta - b$ , and hence  $I(\theta_1, u_g)$  is maximized when  $b = \theta_1$ .

**Exact determinations of best difficulty levels.** It might be of interest to have explicit formulas or numerical tables for the value of  $b$  that minimizes the expression (19.5.12) for  $\beta$ , for given values of  $\theta_1$ ,  $\theta_2$ ,  $\alpha$ , and  $n$ , and given forms of  $P(\theta_1)$  and  $P(\theta_2)$  in which  $b$  is the sole variable parameter. It might be convenient and of interest to combine this derivation with derivations for the similar problem in which  $\theta_1$ ,  $\theta_2$ ,  $\alpha$ ,  $\beta$ , and the forms of  $P(\theta_1)$  and  $P(\theta_2)$  are given, and it is required to determine (1) a value of  $b$  that meets (or exceeds) these requirements with the smallest possible number  $n$  of items and (2) to determine that minimum  $n$ . Such formulas are not now available, but the problems stated are readily solved in specific cases by successive numerical trials.

### Exercises

- 19.1. Let  $P(\theta) = \Psi[(1.7)(0.5)(\theta - b)]$ ,  $\theta_1 = 1$ ,  $\theta_2 = 2$ ,  $\alpha = 0.01$ , and  $n = 49$ . Compute  $\beta$  from (19.5.12) for the cases  $b = 0, 1, 2, 3, 0.5, 1.5, 2.5$ , and, if desired, for other successive trial values as well, to approximate a value of  $b$  that minimizes  $\beta$ .
- 19.2. Using the same specifications, but omitting the given value of  $n$  and adding the requirement that  $\beta = 0.10$ , determine approximately, by use of trial values of  $b$ , a value of  $b$  that minimizes the required value of  $n$ .
- 19.3. Give numerical examples relevant to the questions of robustness indicated in connection with the examples in Section 19.3.

### References and Selected Readings

- BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58-16*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January 1957.
- BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)
- BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)
- BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics*, 1958, **29**, 1284 (abstract). (c)
- BIRNBAUM, A., Statistical theory of tests of a mental ability. *Ibid*, 1285 (abstract). (d)
- BIRNBAUM, A., A unified theory of estimation, I. *Annals of Mathematical Statistics*, 1961, **32**, 112-137. (a)

- BIRNBAUM, A., The theory of statistical inference. New York: Institute of Mathematical Sciences, New York University, 1961. (b) (Mimeographed)
- BIRNBAUM, A., On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 1962, **57**, 269-326.
- BIRNBAUM, A., Likelihood. *International Encyclopedia of the Social Sciences*, 1968. (a)
- BIRNBAUM, A., Concepts of statistical evidence. In S. Morgenbesser, P. Suppes, and M. White (Eds.), *Essays in Honor of Ernest Nagel*. New York: St. Martin's Press, 1968. (b)
- BIRNBAUM, A., and A. E. MAXWELL, Classification procedures based on Bayes' formula. *Applied Statistics*, 1960, **9**, 152-169. Reprinted in L. J. Cronbach and Goldine Gleser, *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press, 1965.
- HARTIGAN, J. A., The asymptotically unbiased prior distribution. *Annals of Mathematical Statistics*, 1965, **36**, 1137-1152.
- LEHMANN, E., *Testing statistical hypotheses*. New York: Wiley, 1959.
- LINDGREN, B. W., *Statistical theory*. New York: Macmillan, 1960-1962.
- LORD, F. M., A theory of test scores. *Psychometric Monograph*, No. 7. Chicago: University of Chicago Press, 1952. (a)
- LORD, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17**, 181-194. (b)
- LORD, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18**, 57-76.
- MOSTELLER, F., and D. WALLACE, *Inference and disputed authorship: The Federalist*. Reading, Mass.: Addison-Wesley, 1964.
- NOVICK, M. R., and W. J. HALL, A Bayesian indifference procedure. *Journal of the American Statistical Association*, 1965, **60**, 1104-1117.
- RAO, C. R., *Linear statistical inference and its applications*. New York: Wiley, 1965.
- SAVAGE, L. J., *The foundations of statistical inference*. New York: Wiley, 1962.
- TUKEY, J., The future of data analysis. *Annals of Mathematical Statistics*, 1962, **33**, 1-67.

# ESTIMATION OF AN ABILITY

## 20.1 Introduction

In this chapter we shall systematically develop some general methods of assessing the usefulness of test models for estimating a subject's ability  $\theta$ . We shall then show that these methods provide general guidance and specific working techniques for selecting items and for designing and constructing tests for specified purposes.

We shall develop these methods in terms of the normal approximation to the distribution of scoring formulas  $x = x(\mathbf{v})$ , and in terms of point and confidence limit estimators  $\theta^*(x, \alpha)$  of  $\theta$ , which we introduced in Sections 17.4 and 17.7. There we saw that the precision properties of estimators based on a given scoring formula are usefully represented by (1) the variance of the scoring formula  $\sigma^2[x(\mathbf{V}), \theta]$ , and (2) the derivative  $\partial\mathcal{E}[x(\mathbf{V}) | \theta]/\partial\theta$ , which specifies how the mean of the scoring formula depends on  $\theta$ .

In particular, we saw that these precision properties are summarized in the *information function of a given scoring formula*

$$I[\theta, x(\mathbf{v})] = \frac{1}{\sigma^2[x(\mathbf{V}), \theta]} \left\{ \frac{\partial}{\partial\theta} \mathcal{E}[x(\mathbf{V}) | \theta] \right\}^2, \quad (20.1.1)$$

where  $x = x(\mathbf{v})$  denotes any given test scoring formula based on any given test model. The model may be represented by its pdf's  $\text{Prob}(\mathbf{V} = \mathbf{v} | \theta)$  or just the cdf's  $F(x | \theta)$  of the given score. For brevity, we shall write  $I[\theta, x(\mathbf{v})] = I(\theta, x)$ .

## 20.2 Some Algebra of Information Functions

If  $x(\mathbf{v})$  has any weighted-sum form

$$x(\mathbf{v}) = \sum_{g=1}^n w_g u_g, \quad (20.2.1)$$

where the  $w_g$  are any positive numbers, then we may substitute (17.7.2) and (17.7.3) in (20.1.1) and obtain

$$I(\theta, x) = \left[ \sum_{g=1}^n w_g^2 P_g(\theta) Q_g(\theta) \right]^{-1} \left[ \sum_{g=1}^n w_g P'_g(\theta) \right]^2, \quad (20.2.2)$$

where

$$P'_g(\theta) = \frac{\partial}{\partial \theta} P_g(\theta).$$

As a case of the score formula  $x(v)$  we may take a single term  $w_g u_g$  of such a score formula or a single item response  $u_g$ , either of which gives the *item information function*

$$I(\theta, u_g) \equiv I(\theta, w_g u_g) = P'_g(\theta)^2 / P_g(\theta) Q_g(\theta). \quad (20.2.3)$$

Now we may easily verify (by application of the Cauchy inequality) that for any given numbers  $C_g, D_g (D_g > 0)$ , we have

$$\left( \sum_{g=1}^n w_g C_g \right)^2 \leq \sum_{g=1}^n (C_g / \sqrt{D_g})^2 \sum_{g=1}^n (w_g \sqrt{D_g})^2, \quad (20.2.4)$$

with equality if and only if  $w_g = AC_g/D_g, g = 1, \dots, n$ , where  $A$  may be any nonzero number. Setting

$$C_g = P'_g(\theta) \quad \text{and} \quad D_g = P_g(\theta) Q_g(\theta) \text{ gives}$$

$$I(\theta, x) \leq \sum_{g=1}^n I(\theta, u_g), \quad (20.2.5)$$

with equality if and only if

$$w_g = P'_g(\theta) / P_g(\theta) Q_g(\theta) = w_g(\theta), \quad g = 1, \dots, n, \quad (20.2.6)$$

except for the usual allowed arbitrary positive constant factor. That is, equality obtains if and only if the  $w_g$  are locally best weights at  $\theta$ . We noted in Section 19.3 that locally best weights independent of  $\theta$  exist only in logistic test models. In general, then,

$$I(\theta, x) < \sum_{g=1}^n I(\theta, u_g),$$

except at those values of  $\theta$ , if any, for which the given weights are locally best; and equality obtains uniformly in  $\theta$  only if the model is logistic (or consists of equivalent items) and if  $x = \sum_{g=1}^n a_g u_g$ . We shall call the right-hand member of (20.2.5),

$$I(\theta) = \sum_{g=1}^n I(\theta, u_g) \equiv \sum_{g=1}^n [P'_g(\theta)]^2 / P_g(\theta) Q_g(\theta), \quad (20.2.7)$$

the *information function of a test*.

We note, therefore, that  $I(\theta)$  is determined by the test model, since it is only the sum of the information functions of its items, and that it is not dependent on

any choice of a score formula  $x(\mathbf{v})$ . Because of these facts and because of the relation (20.2.5), it constitutes an upper bound on each and all of the information functions  $I(\theta, x)$  that may be obtained by the various possible choices of test score formulas of the weighted sum form.

When no score formula of the weighted sum form gives equality for all  $\theta$  in (20.2.5), the precision of estimation represented by such equality is nevertheless usually obtainable by use of statistics of more widely applicable forms such as maximum likelihood estimators, as we shall show in the next section. Also, since the formula (20.2.7) exhibits a basic general additivity property of the information structures of tests in relation to their items, it is a basis for solving a number of problems of appraising and designing of tests, test items, and score formulas and estimators, as we shall show in Sections 20.4 through 20.6.

### 20.3 More General Methods of Estimation: Maximum Likelihood

The relation

$$I(\theta, x) = I(\theta)$$

never holds, uniformly in  $\theta$ , for linear  $x(\mathbf{v})$  except in logistic or equivalent-item test models, as we have indicated in the preceding section. For test models of other forms, the precision of estimation represented by this relation is obtainable in many cases of interest by use of estimators based on certain *nonlinear* score formulas  $x(\mathbf{v})$ . Since each point estimator  $\theta^* = \theta^*(x) = \theta^*[x(\mathbf{v})]$  based on a statistic  $x = x(\mathbf{v})$  is itself a real-valued statistic  $\theta^* = \theta^*[\mathbf{v}]$ , it is appropriate and sometimes convenient to express such precision properties of a given estimator  $\theta^*$  by writing

$$I(\theta, \theta^*[\mathbf{v}]) \equiv I(\theta, \theta^*) = I(\theta), \quad (20.3.1)$$

uniformly in  $\theta$ . [Here  $I(\theta, \theta^*)$  is calculated using the asymptotic mean and variance of  $\theta^*$ , as discussed in Section 17.7.] In many of these cases we shall see that it is convenient to define and use such estimators without reference to any other intermediary statistic  $x(\mathbf{v})$ .

Estimators  $\theta^*$  satisfying (20.3.1) are called *asymptotically efficient*, since the right-hand member of (20.3.1) can be shown to be an upper bound for the information function of any test no matter what estimator is used.\* For our purposes, the most convenient of these estimators are the *maximum likelihood estimators*, customarily denoted by  $\hat{\theta} = \hat{\theta}(\mathbf{v})$ . When  $\mathbf{v}$  is fixed at an observed response pattern, the function  $\text{Prob}(\mathbf{V} = \mathbf{v} | \theta)$  of  $\theta$  is called the *likelihood function*. If, as occurs in most of the models discussed above, this function has, for each  $\mathbf{v}$ , a maximum that is attained at a unique value of  $\theta$ , the *maximum likelihood estimate*  $\hat{\theta} = \hat{\theta}(\mathbf{v})$  is defined to have this value.

---

\* Exceptions, of more theoretical than practical interest, are the *superefficient estimators*. See Kendall and Stuart, 1961, p. 44.

For both theoretical and practical purposes, it is often convenient to avoid writing an explicit formula for  $\hat{\theta}(\mathbf{v})$ , which is in general nonlinear. Provided that the derivative

$$\frac{\partial}{\partial \theta} \log \text{Prob} (\mathbf{V} = \mathbf{v} | \theta)$$

exists for each  $\mathbf{v}$  and is a decreasing function of  $\theta$  that assumes the value 0, the *likelihood equation*

$$\frac{\partial}{\partial \theta} \log \text{Prob} (\mathbf{V} = \mathbf{v} | \theta) = 0 \quad (20.3.2)$$

defines implicitly the maximum likelihood estimator  $\hat{\theta}(\mathbf{v})$ ; for each observed  $\mathbf{v}$ ,  $\hat{\theta} = \hat{\theta}(\mathbf{v})$  can be calculated numerically as the solution of this equation.

Some of the efficiency properties of maximum likelihood estimators can be proved in exact and elementary terms (see Birnbaum, 1961a, pp. 122–127): Let

$$\begin{aligned} \frac{\partial}{\partial \theta} \log \text{Prob} (\mathbf{V} = \mathbf{v} | \theta) &= \frac{\partial}{\partial \theta} \log \prod_{g=1}^n \left[ \frac{P_g(\theta)}{Q_g(\theta)} \right]^{u_g} \prod_{g=1}^n Q_g(\theta) \\ &= \sum_{g=1}^n w_g(\theta) u_g - R(\theta) \\ &= x(\mathbf{v}, \theta) - R(\theta), \end{aligned} \quad (20.3.3)$$

where

$$w_g(\theta) = \frac{\partial}{\partial \theta} \log \frac{P_g(\theta)}{Q_g(\theta)} \quad \text{and} \quad R(\theta) = - \frac{\partial}{\partial \theta} \log \prod_{g=1}^n Q_g(\theta),$$

$w_g(\theta)$  being a locally best weight (see Section 19.3) and  $R(\theta)$  being a nonnegative quantity. We note that  $x(\mathbf{v}, \theta)$  is not a statistic, because of its dependence on  $\theta$ ; but if  $\theta$  is fixed at any chosen value  $\theta'$ , then  $x(\mathbf{v}, \theta')$  is a statistic, and in particular it is the locally best score formula at  $\theta'$ .

One locally best classification rule for abilities in the neighborhood of  $\theta'$  has the form: Classify as high just when  $x(\mathbf{v}, \theta') > R(\theta')$ . Now if  $\mathbf{v}'$  is any response pattern that is classified as high by this rule, we have  $x(\mathbf{v}', \theta') > R(\theta')$ . Under the conditions mentioned above,  $x(\mathbf{v}', \theta)$  is decreasing in  $\theta$ , and hence  $\hat{\theta}(\mathbf{v}') > \theta'$ . Similarly, if  $\mathbf{v}''$  is any response pattern that is classified as low by the rule described, we have  $\hat{\theta}(\mathbf{v}'') \leq \theta'$ . Thus we see that if we compare any maximum likelihood estimate  $\hat{\theta}(\mathbf{v})$  with any specified ability level  $\theta'$  of interest, and if we take  $\hat{\theta}(\mathbf{v}) > \theta'$  as evidence that the true ability value exceeds  $\theta'$  and in similar fashion take  $\hat{\theta}(\mathbf{v}) \leq \theta'$  as evidence that the true ability value does not exceed  $\theta'$ , then we are interpreting maximum likelihood estimates in ways that correspond in formal detail with certain locally best classification rules.

In cases in which the locally best score has approximately a normal distribution at  $\theta'$  (see Section 17.7), the probability  $\text{Prob}[x(\mathbf{V}, \theta') > R(\theta') | \theta']$ , which uniquely characterizes one locally best rule, will be very near  $\frac{1}{2}$ . To

prove this, we note that

$$\begin{aligned}
 \mathcal{E}[x(\mathbf{V}, \theta') | \theta'] &= \sum_{\mathbf{v}} \text{Prob} (\mathbf{V} = \mathbf{v} | \theta') x(\mathbf{v}, \theta') \\
 &= \sum_{\mathbf{v}} \text{Prob} (\mathbf{V} = \mathbf{v} | \theta') \left[ \frac{\partial}{\partial \theta} \log \text{Prob} (\mathbf{V} = \mathbf{v} | \theta) \Big|_{\theta=\theta'} + R(\theta') \right] \\
 &= \sum_{\mathbf{v}} \text{Prob} (\mathbf{V} = \mathbf{v} | \theta') \\
 &\quad \times \left[ \frac{1}{\text{Prob} (\mathbf{V} = \mathbf{v} | \theta')} \frac{\partial}{\partial \theta} \text{Prob} (\mathbf{V} = \mathbf{v} | \theta) \Big|_{\theta=\theta'} \right] + R(\theta') \\
 &= \frac{\partial}{\partial \theta} \left[ \sum_{\mathbf{v}} \text{Prob} (\mathbf{V} = \mathbf{v} | \theta) \right]_{\theta=\theta'} + R(\theta') \\
 &= \frac{\partial}{\partial \theta} 1 + R(\theta') = R(\theta'), \tag{20.3.4}
 \end{aligned}$$

and hence

$$\text{Prob} [x(\mathbf{V}, \theta') \leq R(\theta') | \theta'] \doteq \Phi \left[ \frac{R(\theta') - \mathcal{E}[x(\mathbf{V}, \theta') | \theta']}{\sigma[x(\mathbf{V}, \theta') | \theta']} \right] = \Phi(0) = \frac{1}{2}. \tag{20.3.5}$$

Combining this approximate value with the observation above, that  $x(\mathbf{v}, \theta') > R(\theta')$  for each  $\theta'$  if and only if  $\hat{\theta}(\mathbf{v}) > \theta'$ , we see that under the conditions mentioned, the maximum likelihood estimator is approximately median-unbiased. (The latter property was defined in Section 17.4.)

It can further be shown (Cramér, 1946, p. 500) that *a maximum likelihood estimator has approximately (asymptotically) the normal distribution with mean  $\theta$ , the true ability value, and variance  $1/I(\theta)$ , under conditions satisfied by most of the models under discussion.* Calculating  $I(\theta, \hat{\theta})$  with these asymptotic values, we find that  $I(\theta, \hat{\theta}) = I(\theta)$ , and therefore *the maximum likelihood estimator  $\hat{\theta}$  is asymptotically efficient.*

In those special cases in which  $I(\theta)$  is not dependent on the unknown true value of  $\theta$ , its value  $I$  can be computed and used in formulas for such statistics as

$$\hat{\theta} + \Phi^{-1}(1 - \alpha)/\sqrt{I(\theta)}, \tag{20.3.6}$$

which represents a *maximum likelihood confidence limit estimator* with approximate (asymptotic) confidence coefficient  $1 - \alpha$  (or  $\alpha$ , if  $1 - \alpha < \frac{1}{2}$ ). Such an estimator shares the asymptotic efficiency property of  $\hat{\theta}$  itself. Fortunately, although  $I(\theta)$  varies with  $\theta$  in most cases of interest here, it can be shown (Kendall and Stuart, 1961; Wald, 1942) that replacing the unknown  $\theta$  in  $I(\theta)$  by its estimate  $\hat{\theta}$  and substituting in (20.3.6) gives a *maximum likelihood confidence limit estimator*

$$\hat{\theta}(\mathbf{v}, 1 - \alpha) \equiv \hat{\theta} + \Phi^{-1}(1 - \alpha)/\sqrt{I(\hat{\theta})}, \tag{20.3.7}$$

which also shares the asymptotic efficiency property of  $\hat{\theta}$  under regularity conditions satisfied by most of the models considered here.

### Examples of maximum likelihood estimators of an ability

#### 1. Logistic test model

$$x(\mathbf{v}, \theta) = \sum_{g=1}^n w_g(\theta) u_g, \quad (20.3.8)$$

where

$$\begin{aligned} w_g(\theta) &= \frac{\partial}{\partial \theta} \log \frac{P_g(\theta)}{Q_g(\theta)} = \frac{\partial}{\partial \theta} \log \frac{\Psi[DL_g(\theta)]}{\Psi[-DL_g(\theta)]} \\ &= \frac{\partial}{\partial \theta} \log e^{DL_g(\theta)} = \frac{\partial}{\partial \theta} DL_g(\theta) = Da_g, \end{aligned}$$

and

$$R(\theta) = - \sum_{g=1}^n \frac{\partial}{\partial \theta} \log Q_g(\theta),$$

in which

$$-\frac{\partial}{\partial \theta} \log Q_g(\theta) = -\frac{\partial}{\partial \theta} \log [1 + e^{DL_g(\theta)}]^{-1} = Da_g \Psi[DL_g(\theta)] = Da_g P_g(\theta).$$

Thus the likelihood equation is

$$x(\mathbf{v}, \hat{\theta}) - R(\hat{\theta}) = D \sum_{g=1}^n a_g u_g - D \sum_{g=1}^n a_g P_g(\hat{\theta}) = 0, \quad (20.3.9)$$

or

$$x(\mathbf{v}) \equiv \sum_{g=1}^n a_g u_g = \sum_{g=1}^n a_g P_g(\hat{\theta}) \equiv E \left( \sum_{g=1}^n a_g U_g \mid \hat{\theta} \right),$$

or

$$\sum_{g=1}^n a_g \Psi[DL_g(\hat{\theta})] = \sum_{g=1}^n a_g u_g. \quad (20.3.10)$$

The right-hand member of the equation (20.3.10) is simply a numerical value of the familiar logistic test score, and the left-hand member is a strictly increasing function of  $\hat{\theta}$ . The reader should note that  $\hat{\theta}$  is a one-to-one function of the sufficient statistic  $x(\mathbf{v})$ . No convenient explicit formula for  $\hat{\theta}$  is available for the general logistic test model.

In the special case of equivalent items, the equation becomes

$$P(\hat{\theta}) = \bar{u}, \quad \text{where} \quad \bar{u} = \frac{1}{n} \sum_{g=1}^n u_g, \quad (20.3.11)$$

and where  $P(\theta)$  is the common item characteristic curve. This case admits the explicit solution for the maximum likelihood estimator

$$\hat{\theta} = P^{-1}(\bar{u}) = b + \frac{\Psi^{-1}(\bar{u})}{Da} = b + \frac{1}{Da} \log \left( \frac{\bar{u}}{1 - \bar{u}} \right). \quad (20.3.12)$$

The explicit formula  $\hat{\theta} = P^{-1}(\bar{u})$  also holds for other cases of tests composed

of equivalent items for which the common item characteristic curves have any form  $P(\theta)$  increasing strictly and continuously from zero to one.

The likelihood equation for a general logistic test model can be solved conveniently in the form (20.3.10) by evaluating the left-hand member of the equation (which increases strictly with  $\theta$ ) at trial values of  $\theta$ , using available tables of  $\Psi(t)$  (see Berkson, 1957, pp. 33–34). Alternatively, if maximum likelihood estimation is to be used repeatedly with a given test model, one can prepare a table or graph of the left-hand member of the likelihood function, from which one can read, as accurately as he desires, the maximum likelihood estimate  $\hat{\theta} = \hat{\theta}(x)$  corresponding to each possible value of the score  $x = \sum_{g=1}^n a_g u_g$ . Such a graph corresponds approximately to the contour  $F(x | \theta) = 0.5$  in Fig. 17.4.2, which was used there to determine median-unbiased estimators; this corresponds to the above-mentioned approximate median unbiasedness that holds for maximum likelihood estimators in many cases of interest.

To determine maximum likelihood confidence limits as given by (20.3.7), we must compute

$$I(\hat{\theta}) = D^2 \sum_{g=1}^n a_g^2 \psi[DL_g(\hat{\theta})].$$

This computation can be facilitated by use of available tables of  $\psi(t)$  (Berkson, 1957). Again, for convenience in repeated use, one may prepare a table or graph of the function  $I(\hat{\theta})^{-1/2}$ , the asymptotic estimate of standard error that appears with the maximum likelihood estimate  $\hat{\theta}$  in the formula (20.3.7) for  $\hat{\theta}(v, 1 - \alpha)$ .

*2. Normal ogive model.* Lord (1953, pp. 60–63) has treated this case in some detail. When specialized to the present case of known item parameters, Lord's treatment of maximum likelihood estimation of ability is analogous to the following treatment of the three-parameter logistic model.

*3. Three-parameter logistic model.* The general features of maximum likelihood estimation methods here closely resemble those of the special case in which all guessing probabilities are zero, that is, the case of the general logistic model discussed above. The locally best weights are

$$w_g(\theta) = \frac{\partial}{\partial \theta} \log \frac{P_g(\theta)}{Q_g(\theta)} = Da_g \Psi[DL_g(\theta) - \log c_g],$$

and the likelihood equation can be written as

$$\sum_{g=1}^n a_g \Psi[DL_g(\hat{\theta})] = \sum_{g=1}^n a_g \frac{w_g(\hat{\theta})}{D} u_g.$$

We observe that the left-hand member here is identical to the left-hand member of (20.3.10), the equation for the two-parameter logistic model. The right-hand member varies with  $\hat{\theta}$ . For a fixed response pattern,  $\hat{\theta}$  may be computed numerically by trials, with use of tables of  $\Psi(t)$ .

## 20.4 The Information Functions of Various Test Items

The information functions of items may be regarded as building blocks from which the information function of a test is constructed. This is implied by the basic general additive relation represented in the defining equation (20.2.7):

$$I(\theta) = \sum_{g=1}^n I(\theta, u_g) = \sum_{g=1}^n [P'_g(\theta)]^2 / P_g(\theta) Q_g(\theta). \quad (20.4.1)$$

In this section we shall prepare ourselves to use this relation systematically by analyzing in detail the contributions of individual items to a test information function. For items of several of the test models considered above, we shall determine the item information function  $I(\theta, u_g)$ . We shall also determine three parameters of that function which describe aspects of the item's contribution to the test information function:

1. The maximum of  $I(\theta, u_g)$ ,

$$M_g = \max_{\theta} I(\theta, u_g), \quad (20.4.2)$$

will sometimes be of interest, since at each  $\theta$  we have

$$I(\theta) = \sum_{g=1}^n I(\theta, u_g) \leq \sum_{g=1}^n M_g. \quad (20.4.3)$$

2. We shall denote the value of  $\theta$  at which the maximum is attained by  $\theta_g$ , when this value is uniquely defined.
3. The area under the curve  $I(\theta, u_g)$ ,

$$A_g = \int_{-\infty}^{\infty} I(\theta, u_g) d\theta, \quad (20.4.4)$$

will also be of some interest, since

$$\int_{-\infty}^{\infty} I(\theta) d\theta = \sum_{g=1}^n A_g. \quad (20.4.5)$$

We shall now determine these parameters for specific test models.

1. *Logistic.* For

$$P_g(\theta) = \Psi[DL_g(\theta)] = [1 + \exp D(a_g\theta - b_g)]^{-1},$$

we have

$$P'_g(\theta) = Da_g\Psi'[DL_g(\theta)] \quad \text{and} \quad P_g(\theta)Q_g(\theta) = \psi[DL_g(\theta)]. \quad (20.4.6)$$

Thus the information function of a test item in the logistic model is

$$I(\theta, u_g) = P'_g(\theta)^2 / P_g(\theta)Q_g(\theta) = D^2 a_g^2 \psi[DL_g(\theta)]. \quad (20.4.7)$$

Since the maximum of  $\psi[DL_g(\theta)]$  is  $\psi(0) = \Psi(0)[1 - \Psi(0)] = (\frac{1}{2})^2 = \frac{1}{4}$ , the maximum of  $I(\theta, u_g)$  is

$$M_g = \frac{1}{4}D^2a_g^2 = \frac{1}{4}a_g^2(1.7)^2 = 0.721a_g^2; \quad (20.4.8)$$

it is attained at  $\theta_g = b_g$ . The area under  $I(\theta, u_g)$  is

$$A_g = \int_{-\infty}^{\infty} I(\theta, u_g) d\theta = Da_g \int_{-\infty}^{\infty} Da_g \psi[DL_g(\theta)] d\theta = Da_g = 1.7a_g. \quad (20.4.9)$$

Figure 20.4.1 shows information curves for logistic items with difficulties  $b_g$  all zero, and  $a_g = g$ ,  $g = 1, 2, 3$ . Values  $a_g > 1$  are considered rare in practice; typically, therefore, the area under an item information curve is spread more or less smoothly over at least several units on the  $\theta$  scale rather than concentrated over a small interval of  $\theta$  values (as would be theoretically optimal for some of the classification problems discussed in the previous chapter).

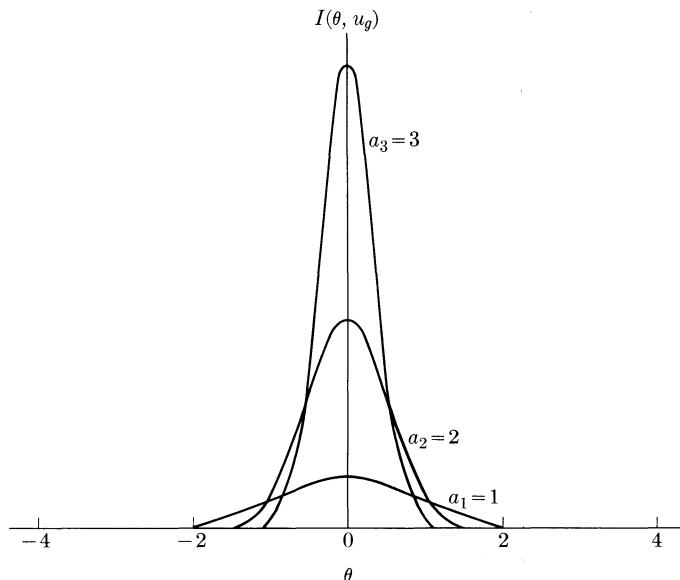


FIG. 20.4.1. Item characteristic curves of the logistic items with  $a_1 = 1$ ,  $a_2 = 2$ ,  $a_3 = 3$ , and  $b_g = 0$ ,  $g = 1, 2, 3$ .

*2. Normal ogive.* We have

$$\begin{aligned} I(\theta, u_g) &= P'_g(\theta)^2/P_g(\theta)Q_g(\theta) \\ &= a_g^2\varphi[L_g(\theta)]^2/\Phi[L_g(\theta)]\Phi[-L_g(\theta)] \end{aligned} \quad (20.4.10)$$

or

$$I(\theta, u_g) = a_g^2J(s), \quad (20.4.11)$$

where

$$J(s) = \varphi(s)^2/\Phi(s)\Phi(-s), \quad \text{and} \quad s = L_g(\theta). \quad (20.4.12)$$

The maximum of  $I(\theta, u_g)$  is attained when  $s = 0$ , that is, at  $\theta_g = b_g$ . The maximum is

$$M_g = a_g^2 / (\sqrt{2\pi})^2 (\tfrac{1}{2})^2 = 2a_g^2 / \pi.$$

The area  $A_g$  could be determined by numerical integration; it is not used below.

*3. Three-parameter logistic.* If we take

$$P_g(\theta) = c_g + (1 - c_g)\Psi[DL_g(\theta)] \quad (20.4.13)$$

(this includes the logistic case above as the special extreme case in which  $c_g = 0$ ), we have

$$P'_g(\theta) = (1 - c_g)Da_g\Psi[DL_g(\theta)] \quad (20.4.14)$$

and

$$P_g(\theta)Q_g(\theta) = (1 - c_g)\{\Psi[DL_g(\theta)] + c_g\Psi[-DL_g(\theta)]^2\}. \quad (20.4.15)$$

Hence the information function of an item in the logistic test model with guessing probabilities is

$$I(\theta, u_g) = (1 - c_g)D^2a_g^2\Psi^2[DL_g(\theta)] / \{\Psi[DL_g(\theta)] + c_g\Psi[-DL_g(\theta)]^2\}. \quad (20.4.16)$$

An alternative form is obtained from (20.4.16) by substituting the relation

$$\begin{aligned} \frac{\psi(t) + c\Psi(-t)^2}{\psi(t)} &= 1 + c \frac{1}{(1 + e^t)} \frac{(1 + e^t)}{e^t} \\ &= 1 + \frac{c}{e^t} = \frac{(e^t/c) + 1}{(e^t/c)} = \frac{e^{t-\log c} + 1}{e^{t-\log c}} \\ &= 1/\Psi(t - \log c). \end{aligned} \quad (20.4.17)$$

This results in

$$I(\theta, u_g) = \{D^2a_g^2(1 - c_g)\Psi[DL_g(\theta)]\} \{\Psi[DL_g(\theta)] - \log c_g\}. \quad (20.4.18)$$

In this form, the information function is expressed as a product in which the first factor is the information function of a corresponding hypothetical logistic item with the same parameters, except that  $c_g = 0$ ; this hypothetical item's characteristic curve is thus the probability of a correct response without guessing in the hypothetical interpretation described in Section 17.3. According to this interpretation, the factor  $(1 - c_g)$  decreases the factor described by exactly the probability  $c_g$  that a subject of any ability who cannot answer correctly without guessing will answer correctly by guessing. The final factor is a cumulative logistic distribution function whose median is  $b_g + (\log c_g)/Da_g$ , which lies below  $b_g$  by an amount proportional to  $\log(1/c_g)$ ; if  $c_g$  is very small, this factor is near unity when  $\theta$  is not far below  $b_g$ . If, for example,  $c_g = 0.2$ , the median is

$$b_g - \frac{(\log_e 5)}{Da_g} = b_g - \frac{1.61}{1.7a_g} = b_g - \frac{0.95}{a_g}. \quad (20.4.19)$$

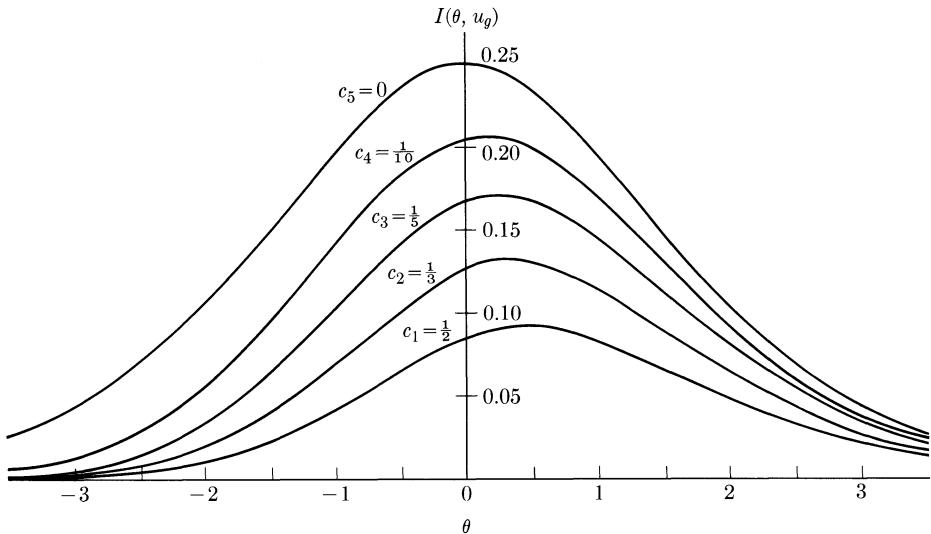


FIG. 20.4.2. Information curves of logistic items with guessing probabilities, with  $a_g = 0.588$ ,  $b_g = 0.0$ ,  $g = 1, \dots, 5$ , and  $c_1 = \frac{1}{2}$ ,  $c_2 = \frac{1}{3}$ ,  $c_3 = \frac{1}{5}$ ,  $c_4 = \frac{1}{10}$ , and  $c_5 = 0$ .

Then if  $a_g = 1/1.7 = 0.588$ ,  $b_g = 0$ , and  $c_g = 0.2$ , we may say that this median is  $-1.61$ .

Another useful form is

$$I(\theta, u_g) = D^2 a_g^2 \psi[DL_g(\theta)] - D^2 a_g^2 P_g(\theta) \psi[DL_g(\theta) - \log c_g]. \quad (20.4.20)$$

Here the first term on the right-hand side is the information function of the hypothetical logistic item referred to, and the second term represents in an additive form the information hypothetically lost because of guessing.

Figure 20.4.2 shows several such information curves for items with  $b_g = 0$  and  $a_g = 1/D = 1/1.7 = 0.588$ , with  $c_g$  taking the values  $\frac{1}{2}$ ,  $\frac{1}{3}$ ,  $\frac{1}{5}$ , and  $\frac{1}{10}$ , respectively. The limiting case  $c_g = 0$  is included for comparison.

Later in this section we shall show that the value  $\theta_g$  of  $\theta$  that maximizes  $I(\theta, u_g)$  is

$$\theta_g = b_g + \frac{1}{D a_g} \log \frac{1 + (1 + 8c_g)^{1/2}}{2}. \quad (20.4.21)$$

Since  $0 < c_g < 1$ , we have

$$0 < \log \frac{1 + (1 + 8c_g)^{1/2}}{2} < \log_e 2 = 0.69.$$

Figure 20.4.2 illustrates the increase of  $\theta_g$  with  $c_g$  for items with  $a_g = 1/D$  and  $b_g = 0$ .

If any number  $n$  of items were available with given common values of  $a_g$  and  $c_g$  and with the values  $b_g$ ,  $g = 1, \dots, n$ , subject to our specification, then, according to (20.4.21), we should specify the common value

$$b_g = \theta' - \frac{1}{Da_g} \log \frac{1 + (1 + 8c_g)^{1/2}}{2}, \quad g = 1, \dots, n, \quad (20.4.22)$$

to maximize

$$I(\theta) = \sum_{i=1}^n I(\theta, u_g)$$

at a given ability level  $\theta'$ . Conclusions very similar to these, quantitatively as well as qualitatively, have been obtained for analogous efficiency problems in the normal ogive test model with guessing probabilities, under the assumption that abilities have a given probability distribution (Lord, 1953, pp. 67–69, and references therein).

To derive the formula (20.4.21) for  $\theta_g$ , we write  $t = DL_g(\theta) = Da_g(\theta - b_g)$  and maximize  $I(\theta, u_g)$  with respect to  $t$ . From (20.4.18), we find that

$$\begin{aligned} \frac{\partial}{\partial t} \log I(\theta, u_g) &= \frac{\partial}{\partial t} [\log \psi(t) + \log \Psi(t - \log c_g)] \\ &= 2\Psi(-t) - 1 + \Psi(-t + \log c_g) \\ &= \frac{2}{1 + e^t} - 1 + \frac{1}{1 + e^t/c_g} = \frac{1 - e^t}{1 + e^t} + \frac{c_g}{c_g + e^t} \\ &= (2c_g + e^t - e^{2t})/(c_g + e^t)(1 + e^t). \end{aligned} \quad (20.4.23)$$

Setting this derivative equal to zero and solving for  $t$  yields

$$t \equiv Da_g(\theta_g - b_g) = \log(\frac{1}{2} + \frac{1}{2}\sqrt{1 + 8c_g}), \quad (20.4.24)$$

which gives the stated formula for  $\theta_g$ .

For the area  $A_g$  under an item information curve, we have

$$\begin{aligned} A_g &= \int_{-\infty}^{\infty} I(\theta, u_g) d\theta \\ &= D^2 a_g^2 (1 - c_g) \int_{-\infty}^{\infty} \Psi[DL_g(\theta)] \Psi[DL_g(\theta) - \log c_g] d\theta. \end{aligned} \quad (20.4.25)$$

After making the substitution  $r = 1 + \exp DL_g(\theta)$  and integrating by the method of partial fractions, we find that

$$A_g = Da_g \frac{c_g \log c_g + 1 - c_g}{(1 - c_g)}. \quad (20.4.26)$$

As  $c_g \rightarrow 0$ , the right-hand member increases to  $Da_g$ , the exact value (Eq. 20.4.9) of  $A_g$  found for two-parameter logistic items.

## 20.5 The Information Functions of Various Tests

In the present section we shall describe some examples of test information curves as they are related to item information curves by the additive definition (20.2.7):

$$I(\theta) = \sum_{g=1}^n I(\theta, u_g).$$

The examples will also illustrate the discussion of problems of test design and item selection in the following section.

*1. Tests with equivalent items.* If  $P_1(\theta)$  is the common ICC of the  $n$  items constituting a test, and if  $I(\theta, u_1)$  is the common item information function determined by  $P_1(\theta)$ , then the information function of the test is

$$I(\theta) = nI(\theta, u_1). \quad (20.5.1)$$

If  $M_1$  denotes the maximum of  $I(\theta, u_1)$ , and that maximum is attained at  $\theta = \theta_1$ , then the maximum of  $I(\theta)$  is  $nM_1$ , and is also attained at  $\theta = \theta_1$ . If  $A_1$  is the area under  $I(\theta, u_1)$ , then the area under  $I(\theta)$  is  $nA_1$ . Each detail of an item information function thus determines in a simple way a corresponding detail of the information function of a test consisting of any number  $n$  of such equivalent items.

*2. Tests with nonequivalent items.* A hypothetical test consisting of only several logistic items, with  $a_g$ -values appreciably above the range encountered in practice, and with unequal  $b_g$ -values, would have an information function very near zero except for a very high peak near each  $b_g$ -value. This configuration reflects the function's high discriminating power in the neighborhood of each  $b_g$ -value but otherwise low precision. We discussed such a hypothetical test in Section 16.5, where we noted that items with extremely high  $a_g$ -values, if available in practice, would be ideal for classification into ordered intervals of ability levels, but that they would be of limited value for estimation of ability with good precision over an appreciably wide range of  $\theta$ -values. In a somewhat different theoretical context (where some probability distribution of abilities is assumed), there is a logical possibility that a sufficiently extreme increase of  $a_g$ -values, while other parameters remain fixed, could result in lowering the precision of estimation obtainable with a given test. This possibility has been recognized and discussed by Tucker (1946), Loevinger (1954, as "The Attenuation Paradox in Test Theory"), Solomon (1956), and Sitgreaves (1961). A number of authors have suggested that this "paradox" can be resolved or avoided by use of items with suitably spaced item difficulty parameters  $b_g$ .

Clearly, if each  $a_g$  is extremely high, then one has virtually error-free discrimination between ability levels above and below each  $b_g$ -value represented in the test, and virtually no other information for discrimination or estimation. However, in the cases encountered in practice,  $a_g$ -values exceeding unity are rare; the total area  $1.7a_g$  under the item information function typically is spread rather smoothly over at least several units on the  $\theta$  scale, with only

moderate concentration around  $b_g$ . Thus the greatest concentration practically attainable of the area under a test information curve is obtained by taking items with some common difficulty level, along with  $a_g$ -values as high as possible; but this gives at best only the same limited degree of concentration found in the information curves of the individual items with the highest  $a_g$ -values. *Thus test information curves in practice tend to be less concentrated than item information curves, by an amount that depends primarily on the variation among item difficulty parameters.*

A rough but useful concept here is that the area under the information curve tends to be a smoothed version of the histogram formed by adding, for each item in a logistic test model, a square of area  $Da_g$ , centered at  $\theta = b_g$ . Some-what refined examples of the use of this concept are given next.

*3. Uniformly distributed item difficulties.* Because there is some tendency in practice for discrimination parameters  $a_g$  in normal ogive or logistic models to have a limited range, as mentioned above, while item difficulty parameters  $b_g$  tend to vary appreciably, it is of interest to consider test models with common  $a_g$ -values but with unequal  $b_g$ -values distributed in certain regular patterns. Such oversimplified, schematized versions of cases encountered in practice can provide useful insights, approximations, and guidance.

Consider a logistic test model with  $n$  items having discrimination parameters with the common value  $a_1$ , and with difficulty parameters  $b_g$  uniformly spaced, so that, for some  $\epsilon > 0$ ,

$$b_2 = b_1 + \epsilon, \quad b_3 = b_1 + 2\epsilon, \quad \dots, \quad b_n = b_1 + (n - 1)\epsilon.$$

If  $n$  is not small and  $a_1$  is not large, and if the range of difficulty parameters  $(n - 1)\epsilon$  is not narrow, then the continuous uniform probability density function

$$q(b) = \begin{cases} 1/n\epsilon & \text{for } b_1 - \epsilon/2 \leq b \leq b_n + \epsilon/2, \\ 0 & \text{for other values of } b, \end{cases} \quad (20.5.2)$$

will give a rough approximation to the formal discrete distribution function of the  $b_g$ -values, which will, in turn, give useful approximation formulas for  $I(\theta)$ . We have

$$\begin{aligned} I(\theta) &= D^2 a_1^2 \sum_{g=1}^n \psi[Da_1(\theta - b_g)] \doteq D^2 a_1^2 n \int_{b_1 - \epsilon/2}^{b_n + \epsilon/2} \psi[Da_1(\theta - b)] q(b) db \\ &= \frac{Da_1}{\epsilon} \int_{b_1 - \epsilon/2}^{b_n + \epsilon/2} Da_1 \psi[Da_1(\theta - b)] db \\ &= \frac{Da_1}{\epsilon} \left\{ \Psi \left[ Da_1 \left( b_n + \frac{\epsilon}{2} - \theta \right) \right] - \Psi \left[ Da_1 \left( b_1 - \frac{\epsilon}{2} - \theta \right) \right] \right\} \\ &\doteq \frac{1.7a_1}{\epsilon} \left\{ \Phi \left[ a_1 \left( b_n + \frac{\epsilon}{2} - \theta \right) \right] - \Phi \left[ a_1 \left( b_1 - \frac{\epsilon}{2} - \theta \right) \right] \right\}. \end{aligned} \quad (20.5.3)$$

Thus, for  $\theta$  well within the range of  $b_g$ -values, under the conditions indicated, the last factor will be near unity. Therefore

$$I(\theta) \doteq \frac{1.7a_1}{\epsilon}, \quad \text{for } b_1 + \frac{2}{a_1} < \theta < b_n - \frac{2}{a_1}; \quad (20.5.4)$$

that is,  $I(\theta)$  will have approximately the constant value indicated, which is proportional to the number  $(1/\epsilon)$  of items located (by their difficulty levels  $b_g$ ) on each unit of the  $\theta$  scale and to the area  $1.7a_1$  under an individual item information curve.

*4. Normally distributed item difficulties.* Evidently there is some tendency toward limited concentration of the difficulty parameters  $b_g$  in cases encountered in practice. Let us consider a schematized example that represents this feature. Suppose that a test's  $n$  items have the identical parameters  $a_g \equiv a_1, g = 1, \dots, n$ , and that the item difficulty parameters  $b_g, g = 1, \dots, n$ , are spread approximately in the form of a normal distribution with mean and variance

$$\bar{b} = \frac{1}{n} \sum_{g=1}^n b_g \quad \text{and} \quad \sigma_b^2 = \frac{1}{n} \sum_{g=1}^n (b_g - \bar{b})^2,$$

respectively. This normal distribution is represented by the pdf

$$q(b) = \frac{1}{\sigma_b} \varphi\left(\frac{b - \bar{b}}{\sigma_b}\right). \quad (20.5.5)$$

Both Lawley (1943) and Lord (1952, 1953) have considered this case for the normal ogive model. For the logistic model, we have

$$\begin{aligned} I(\theta) &= D^2 a_1^2 \sum_{g=1}^n \psi[D a_1(\theta - b_g)] \doteq D^2 a_1^2 n \int_{-\infty}^{\infty} \psi[D a_1(\theta - b)] q(b) db \\ &= D^2 a_1^2 \frac{n}{\sigma_b} \int_{-\infty}^{\infty} \psi[D a_1(\theta - b)] \varphi\left(\frac{b - \bar{b}}{\sigma_b}\right) db. \end{aligned} \quad (20.5.6)$$

An approximation that plausibly is close here is the one obtained by replacing the logistic by a normal density function:

$$\begin{aligned} I(\theta) &\doteq D a_1 \frac{n}{\sigma_b} \int_{-\infty}^{\infty} a_1 \varphi[a_1(\theta - b)] \varphi\left(\frac{b - \bar{b}}{\sigma_b}\right) db \\ &= D a_1^2 \frac{n}{2\pi\sigma_b} \int_{-\infty}^{\infty} \exp[-\frac{1}{2}a_1^2(\theta - b)^2] \exp\left[-\frac{1}{2\sigma_b^2}(b - \bar{b})^2\right] db \\ &= D a_1 \frac{n}{h} \varphi\left(\frac{\theta - \bar{b}}{h}\right), \quad \text{where } h^2 = \sigma_b^2 + \frac{1}{a_1^2}. \end{aligned} \quad (20.5.7)$$

That is, when a test is constituted by an appreciable number  $n$  of items with common discrimination parameter  $a_1$  and with various difficulty parameters  $b_g$

having a distribution that formally approximates a normal distribution with mean  $\bar{b}$  and variance  $\sigma_b^2$ , the test information curve has approximately the form of a normal pdf, with mean  $\bar{b}$  and variance  $h^2 = \sigma_b^2 + 1/a_1^2$ , multiplied by the constant  $nDa_1$ .

## 20.6 Problems of Test Design and Item Selection

In models considered in practice, we usually have  $30 \leq n \leq 100$ . Also, if abilities are scaled so that their distribution is normal ( $\mu = 0$ ,  $\sigma^2 = 1$ ) over the population of interest, then we usually have  $a_g < 2$ , where  $a_g$  is the discrimination parameter of a logistic item with characteristic curve  $\Psi[1.7a_g(\theta - b_g)]$ . The difficulty parameters  $b_g$  usually vary over the range  $-3 < b_g < 3$ ; however, when  $a_g$  is close to zero, the  $b_g$  may greatly exceed 3 in absolute value. For such models, the approximations described in the preceding section tend to be close. In many cases, the information function of such a model will be approximately equal to those of several other such models, as we shall illustrate.

If two test models have approximately equal information functions  $I_1(\theta)$  and  $I_2(\theta)$ , respectively, then the test models are approximately equivalent, in terms of those statistical precision properties of estimators and classification rules characterized approximately by information functions. It follows that problems of designing a test model that has an approximately prescribed precision, in the sense of a prescribed information function, typically have non-unique solutions; this allows choice among indicated alternative models, based on such considerations as convenience, availability of items of various types, and length of tests.

For individual logistic items, the range indicated above for  $n$  and  $a_g$  entails the bounds

$$(1.7)(0.2) < A_g < (1.7)(1), \quad \text{or} \quad 0.34 < A_g < 1.7, \quad (20.6.1)$$

on the area  $A_g$  under an item information function, and the bounds

$$M_g \equiv \frac{1}{4}(1.7)^2 a_g^2 < \frac{1}{4}(1.7)^2(1), \quad \text{or} \quad M_g < 0.72, \quad (20.6.2)$$

on the maximum  $M_g$  of an item information function. For test information functions, this range entails, as rather wide bounds,

$$(30)(0.34) < \sum_g A_g < 100(1.7), \quad \text{or} \quad 10 < \sum_g A_g < 170, \quad (20.6.3)$$

and

$$I(\theta) < 100(0.72), \quad \text{or} \quad I(\theta) < 72, \quad \text{for all } \theta. \quad (20.6.4)$$

The following examples illustrate the relevance of such bounds and inequalities for problems of design of test models.

1. Consider the requirement that the asymptotic variance of the maximum likelihood estimator of  $\theta$ , on the basis of a logistic test model, be no greater

than  $0.04 = (0.2)^2$  in the neighborhood of  $\theta = 1$ ; that is,

$$1/I(1) \leq 0.04, \quad \text{or} \quad I(1) \geq 25.$$

This entails  $\sum_{g=1}^n M_g \geq 25$ ; this, in combination with the above bound  $M_g < 0.72$ , entails  $n > 25/0.72$ , or  $n \geq 35$ . The requirement could be met by a logistic model with about  $n = 35$  items, provided that all the item difficulty parameters  $b_g$  were concentrated very near  $\theta = 1$  and all the item discrimination parameters  $a_g$  were very nearly as high as unity.

2. Consider the requirement that the precision indicated in (1) be attained with a logistic test model not only at  $\theta = 1$ , but for  $0 < \theta < 3$ ; that is,  $I(\theta) \geq 25$  for  $0 < \theta < 3$ . This entails

$$\int_0^3 I(\theta) d\theta \geq 75 \quad \text{or} \quad \sum_{g=1}^n A_g \geq 75.$$

This, in combination with the bound  $A_g < 1.7$ , entails  $n > 75/1.7$ , or  $n \geq 45$ . This area requirement could be met by using about 45 items, if items with  $a_g$ -values as high as unity were available.

The problem of designing a logistic model whose information curve meets certain given requirements can usefully be regarded as a problem of filling the area under a target information curve by adding contributions of areas  $A_g = 1.7a_g$  from respective items, each more or less concentrated about a respective difficulty level  $b_g$ . From this standpoint, each of the following groups of logistic items would contribute the same total area toward  $\sum_{g=1}^n A_g$ , namely, 5.1 units of area:

30	items with $a_g = 0.10$ ,	5	items with $a_g = 0.60$ ,
21	items with $a_g = 0.14$ ,	3	items with $a_g = 1.00$ ,
15	items with $a_g = 0.20$ ,	1	item with $a_g = 3.00$ .
9	items with $a_g = 0.33$ ,		

Such a scale for appraising the relative values of items requires the qualification that it may exaggerate the usefulness of items with low  $a_g$  values, in the following respect: If  $a_g \leq 0.33$ , then more than 30% of the area under a logistic item information curve lies outside the interval  $-3 < \theta < 3$ , to which interest is usually restricted, and so will not contribute to filling the area under a target information curve in problems such as (2) above. Again, if  $a_g \leq 0.2$ , more than 55% of the area falls outside the interval  $-3 < \theta < 3$ . On the other hand, for  $a_g$  near unity, if  $b_g$  satisfies  $-1 < b_g < 1$ , then less than 5% of the area falls outside the interval  $-3 < \theta < 3$ .

3. Consider that we specify a target test information function having the form of a given normal pdf with mean  $\mu$  and standard deviation  $\sigma$ , multiplied by

a positive constant  $C$  appreciably larger than unity. We require one or more logistic test models with information curves satisfying

$$I(\theta) \geq \frac{C}{\sigma} \varphi \left( \frac{\theta - \mu}{\sigma} \right),$$

preferably with approximate equality. To fill the area  $C$  under the target curve using logistic items with  $a_g < 1$ , more than  $C/(1.7)(1) = C/1.7$  items are required; using items with  $a_g < 0.5$ , more than  $C/(1.7)(0.5) = C/0.85$  items are required.

If the prescribed value of  $\sigma$  is such that  $1/\sigma$  is exceeded by the  $a_g$ -values of available items, we may take  $n \doteq C/Da_1$  items with  $a_g = a_1, g = 1, \dots, n$ , and with  $b_g$ -values varying approximately according to a normal distribution, with mean and variance

$$\bar{b} \equiv \frac{1}{n} \sum_{g=1}^n b_g = \mu \quad \text{and} \quad \sigma_b^2 \equiv \frac{1}{n} \sum_{g=1}^n (b_g - \bar{b})^2 = \sigma^2 - \frac{1}{a_1^2},$$

respectively. Applying the result of example (4) of Section 20.5, we see that this gives a test model with information curve approximately equal to the target. The nonuniqueness of solutions of test design problems mentioned above is illustrated here by the fact that we may take items with any common value  $a_g \equiv a_1$ , subject to  $a_1 > 1/\sigma$ .

By taking  $n \doteq C$  equivalent logistic items with  $a_g = 1/\sigma$  and  $b_g = \mu$ ,  $g = 1, \dots, n$ , we would obtain a test with

$$I(\theta) = nDa_1\psi[Da_1(\theta - b_1)] \doteq \frac{C}{\sigma} \varphi \left( \frac{\theta - \mu}{\sigma} \right),$$

where the last approximation is as rough as that of a logistic to a normal density function. The last approximation can be improved somewhat by introducing some variation among the  $b_g$ , as discussed above. If the prescribed value  $\sigma$  is such that  $1/\sigma$  appreciably exceeds the values  $a_g$  of the logistic items available, we see that no close approximation to the target curve is possible; by taking enough items with  $b_g = \mu$  to equal the target curve at  $\theta = \mu$  so that

$$\sum_{g=1}^n Da_g\psi(0) = \frac{D}{4} \sum_{g=1}^n a_g \geq \frac{C}{\sigma} \varphi(0) = \frac{C}{\sigma\sqrt{2\pi}},$$

we necessarily obtain a test model whose information curve appreciably exceeds the target for  $\theta$  not near  $\mu$ .

4. Consider a requirement of precision expressed in terms of a function  $G(\theta)$  which is to be estimated. Let  $I[G] \equiv I[G(\theta)]$  denote the information function of any given test, any given increasing differentiable function  $G = G(\theta)$

being taken as the unknown parameter of interest. For example, we might represent the quantile-score ranking of an individual with ability  $\theta$  in a population where abilities have a standard normal distribution by the function  $G(\theta) = \Phi(\theta)$ .

We have the identity

$$I[G] \equiv \frac{I(\theta)}{[\partial G(\theta)/\partial\theta]^2}.$$

Taking  $G(\theta) = \Phi(\theta)$  gives  $I[G] = I(\theta)/\varphi(\theta)^2$  or  $I(\theta) = I[G]\varphi(\theta)^2$ . By the last relation, any requirement of precision of estimation of  $G$ , stated in terms of a target information function  $I^*[G]$  for  $I[G]$ , can be restated in terms of a target information function  $I^*(\theta)$  for  $I(\theta)$ , namely,  $I^*(\theta) = I^*[G]\varphi(\theta)^2$ . This form, with  $\theta$  as argument, is convenient if item information functions and their properties happen to be available in a form with  $\theta$  rather than  $G$  as argument, as they are in our discussions above.

If we take  $I^*[G] \equiv K^2$ , where  $K$  is any positive constant, the requirement is: The estimation of  $G$  with asymptotic variance not to exceed  $1/K^2$ . This gives

$$\begin{aligned} I^*(\theta) &= K^2 \varphi(\theta)^2 = \frac{K^2}{2\pi} (e^{-\theta^2/2})^2 = \frac{K^2}{2\pi} e^{-\theta^2} = \frac{K^2}{2\sqrt{\pi}} \frac{\sqrt{2}}{\sqrt{2\pi}} e^{-(\sqrt{2}\theta)^2/2} \\ &= \frac{K^2}{2\sqrt{\pi}} \sqrt{2} \varphi(\sqrt{2}\theta). \end{aligned}$$

The latter is a case of the above example (3), with  $C = K^2/2\sqrt{\pi}$ ,  $\mu = 0$ , and  $\sigma = 1/\sqrt{2}$ ; and our discussion of that example applies also to the present one. If  $K = 0.02$ , the requirement is: The estimation of an examinee's percentile ranking in such a population, with standard error not to exceed 2%. We note again that the techniques of estimation described here, and throughout Chapters 17 through 20, do not depend on assumptions and interpretations in terms of a population distribution of abilities; and that when in fact such assumptions hold with respect to examinees whose abilities are to be estimated, then it is more appropriate and efficient to use estimation methods of the kind introduced by Birnbaum (1967).

## 20.7 Relative Precisions or Efficiencies of Various Test Designs, Test Score Formulas, and Estimators

Let

$$I_1(\theta, x_1) \equiv I_1[\theta, x_1(\mathbf{v})] \quad \text{and} \quad I_2(\theta, x_2) \equiv I_2[\theta, x_2(\mathbf{v})],$$

respectively, denote the information functions of any two test models and respective score formulas (or estimators)  $x_i \equiv x_i(\mathbf{v})$ ,  $i = 1, 2$ . The initial subscripts refer to the respective test models and may be deleted when two possible score formulas or estimators  $x_i(\mathbf{v})$ ,  $i = 1, 2$ , are considered in connection with

a single test model. In that case, the ratio

$$\text{RE}(\theta, x_1, x_2) = I(\theta, x_1)/I(\theta, x_2) \quad (20.7.1)$$

is called the *relative efficiency* (at  $\theta$ ) of  $x_1$  to  $x_2$ . This function represents the relative precisions of estimators based on  $x_1$  and  $x_2$ , respectively, being just the reciprocal of the ratio of the respective asymptotic variances at  $\theta$ .

In the special cases where  $x_2$  is such that  $I(\theta, x_2) = I(\theta)$  for all  $\theta$ , the ratio

$$\text{Eff}(\theta, x_1) = I(\theta, x_1)/I(\theta) \equiv I(\theta, x_1)/I(\theta, \hat{\theta}) \equiv \text{RE}(\theta, x_1, \hat{\theta}) \quad (20.7.2)$$

is called simply the *efficiency* (at  $\theta$ ) of  $x_1$ . We recall that  $I(\theta, x_2) = I(\theta)$  when  $x_2(v) = \hat{\theta}(v)$ , the maximum likelihood estimator, or when  $x_2 = \sum_{g=1}^n a_g u_g$  in a logistic test model.

An estimator  $\theta^* = \theta^*(v)$ , for which  $\text{Eff}(\theta, \theta^*) = 1$  for all  $\theta$ , is called *efficient*. Examples are the maximum likelihood estimators under conditions indicated in Section 20.3 (or weaker conditions: see Cramér, 1946, pp. 498–506).

Returning to the general case of two possible test models for use in connection with a given latent trait  $\theta$ , we now restrict consideration to one asymptotically efficient estimator based on each model, say  $\hat{\theta}_1$  and  $\hat{\theta}_2$ , respectively. Then the ratio

$$\text{RP}(\theta) = I_1(\theta)/I_2(\theta) \equiv I_1(\theta, \hat{\theta}_1)/I_2(\theta, \hat{\theta}_2) \quad (20.7.3)$$

will be called the *relative precision* (at  $\theta$ ) of the given respective test models. This is just the ratio of asymptotic variances of maximum likelihood estimators  $\sigma_{\hat{\theta}_2}^2/\sigma_{\hat{\theta}_1}^2$  at  $\theta$ , since  $\hat{\theta}_i$  is efficient and since therefore  $\sigma_{\hat{\theta}_i}^2 = 1/I(\theta)$ .

Clearly  $\text{RE}(\theta, x_1, x_2)$ , the relative efficiency of two score formulas or estimators in one test model, and  $\text{RP}(\theta)$ , the relative precision of two test models, are special cases of the general *relative precision function*

$$\text{RP}(\theta) = I_1(\theta, x_1)/I_2(\theta, x_2), \quad (20.7.4)$$

which represents the relative precisions obtainable with any two estimators or scores  $x_1$  and  $x_2$ , based, usually, on different test models.

Many examples are provided by the various test models, test designs, score formulas, and estimators discussed above. The following section is devoted to detailed consideration of the efficiency of the simple unweighted score formula in the general logistic test model.

## 20.8 Efficiency of Unweighted Scores in the Logistic Model

Using (20.2.7) and (20.4.7), we have the information function

$$I(\theta) = D^2 \sum_{g=1}^n a_g^2 \psi[DL_g(\theta)] \quad (20.8.1)$$

for any given logistic test model; and for  $x(\mathbf{v}) = \sum_{g=1}^n a_g u_g$ , we have  $I(\theta, x) \equiv I(\theta, \bar{u}) \equiv I(\theta)$ , for all  $\theta$ . In the same model, for the unweighted score formula

$$\bar{u} = \frac{1}{n} \sum_{g=1}^n u_g,$$

we have

$$I(\theta, \bar{u}) \equiv \frac{[\partial \mathcal{E}(\bar{U} | \theta) / \partial \theta]^2}{\sigma^2(\bar{U}, \theta)} = \frac{\left\{ D \sum_{g=1}^n a_g \psi[DL_g(\theta)] \right\}^2}{\sum_{g=1}^n \psi[DL_g(\theta)]}, \quad (20.8.2)$$

as we have shown in Section 17.7. Hence the efficiency of the unweighted score formula  $\bar{u}$  is

$$\begin{aligned} \text{Eff}(\theta, \bar{u}) &= \frac{I(\theta, \bar{u})}{I(\theta)} = \frac{\left\{ \sum_{g=1}^n a_g \psi[DL_g(\theta)] \right\}^2}{\left\{ \sum_{g=1}^n \psi[DL_g(\theta)] \right\} \left\{ \sum_{g=1}^n a_g^2 \psi[DL_g(\theta)] \right\}} \\ &= \frac{\left( \sum_g a_g \psi_g \right)^2}{\left( \sum_g \psi_g \right) \left( \sum_g a_g^2 \psi_g \right)}. \end{aligned} \quad (20.8.3)$$

In the general case of unequal  $a_g$ , this function is less than unity for each  $\theta$ .

To examine this function in some quantitative detail, we specialize our consideration to simple cases sharing some features with cases occurring in practice. Consider the special case of equal item difficulties, say  $b_g \equiv b_1$ ,  $g = 1, \dots, n$ , at ability level  $\theta = b_1$ . Since  $\psi(0) = \frac{1}{4}$ , we have

$$\text{Eff}(b_1, \bar{u}) = \frac{\left( \sum_{g=1}^n a_g \right)^2}{n \left( \sum_{g=1}^n a_g^2 \right)} = \frac{\left( (1/n) \sum_{g=1}^n a_g \right)^2}{(1/n) \sum_{g=1}^n a_g^2}. \quad (20.8.4)$$

Writing

$$\bar{a} = \frac{1}{n} \sum_{g=1}^n a_g, \quad \bar{a}^2 = \frac{1}{n} \sum_{g=1}^n a_g^2, \quad \sigma_a^2 = \frac{1}{n} \sum_{g=1}^n (a_g - \bar{a})^2 \equiv \bar{a}^2 - \bar{a}^2,$$

$$\text{CV}(a) \equiv \text{coefficient of variation among the } a_g = \sigma_a / \bar{a}, \quad (20.8.5)$$

we have

$$\text{Eff}(b_1, \bar{u}) = \frac{\bar{a}_g^2}{\bar{a}^2} = \frac{\bar{a}_g^2}{\bar{a}^2 + \sigma_a^2} = \frac{1}{1 + \sigma_a^2 / \bar{a}^2} = \frac{1}{1 + \text{CV}(a)^2}. \quad (20.8.6)$$

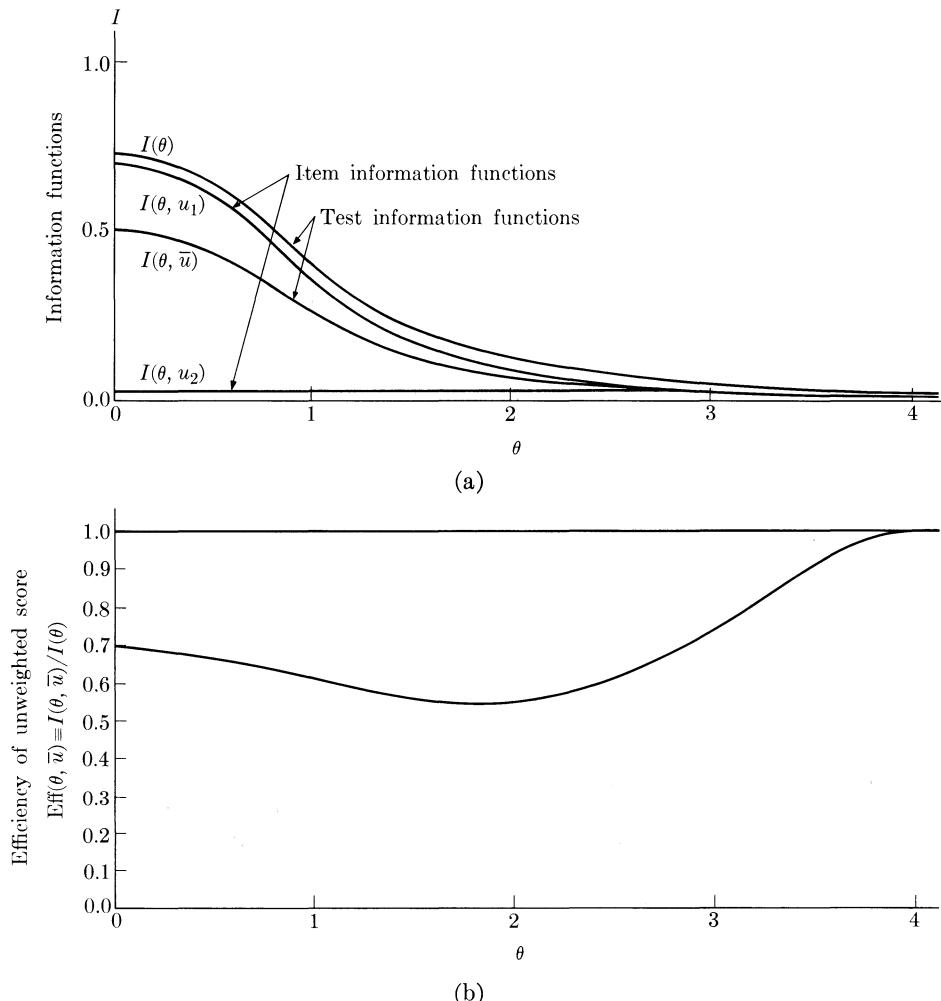


FIG. 20.8.1. Part (a) shows two item information functions,  $I(\theta, \mu_1)$  and  $I(\theta, \mu_2)$ . It also shows two test information functions:  $I(\theta)$ , with optimal weighted score formula ( $x = 0.98\mu_1 + 0.204\mu_2$ ), and  $I(\theta, \bar{\mu})$ , with unweighted formula ( $x = \bar{\mu}$ ).

The function  $\text{Eff}(\theta, \bar{u})$  is graphed in Figs. 20.8.1 through 20.8.4 for some cases in which  $b_g = 0$  for all items. Figure 20.8.1a shows the information curves  $I(\theta, u_1)$  and  $I(\theta, u_2)$  of two items, with  $a_1 = 0.7/\sqrt{0.51} \doteq 0.980$  and  $a_2 = 0.2/\sqrt{0.96} \doteq 0.204$ , respectively; these values represent extremes of the range of  $a_g$  encountered in practice. For comparison, the figure also includes the information curve  $I(\theta) \equiv I(\theta, \hat{\theta}) \equiv I(\theta, x)$  of the hypothetical logistic test model containing just these two items, in which  $x(v) = a_1u_1 + a_2u_2 = 0.98u_1 + 0.204u_2$ . By multiplying this function  $I(\theta)$  by  $n/2$ , we obtain the information

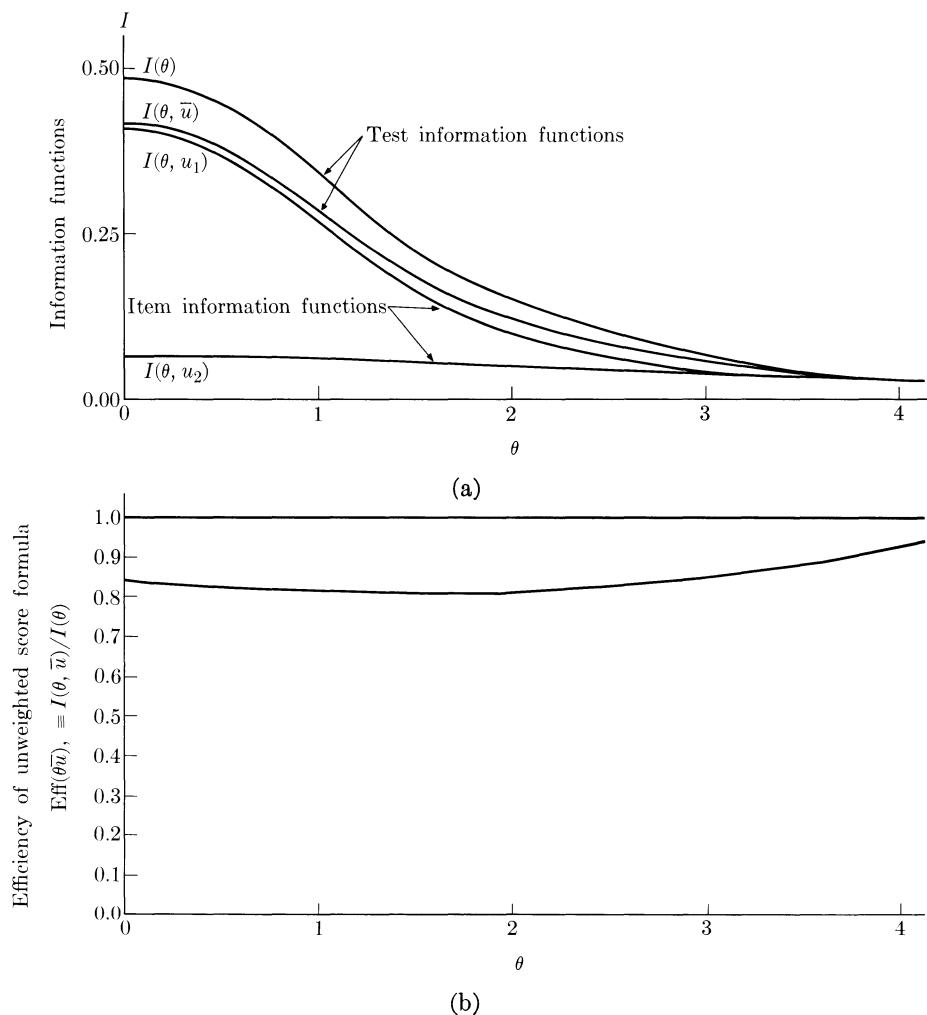


FIG. 20.8.2. Part (a) shows the two item information functions  $I(\theta, \mu_1)$  and  $I(\theta, \mu_2)$ . It also shows the two test information functions  $I(\theta)$  and  $I(\theta, \bar{u})$ , with unweighted and optimal weighted score formulas ( $x = 0.75\mu_1 + 0.315\mu_2$ ).

function of a logistic test model consisting of  $n$  items, of which half have each of the extreme forms indicated. We see that the addition of item 2 to item 1 provides a small but appreciable gain, which is realized if the optimal weighted score formula is used.

For further comparison, the figure includes the information function  $I(\theta, \bar{u})$  of the unweighted score based on the same hypothetical test of two items. When the latter is multiplied by  $n/2$ , we have the information function of  $\bar{u}$  in a logistic test model with  $n$  items, half having each of the forms indicated.

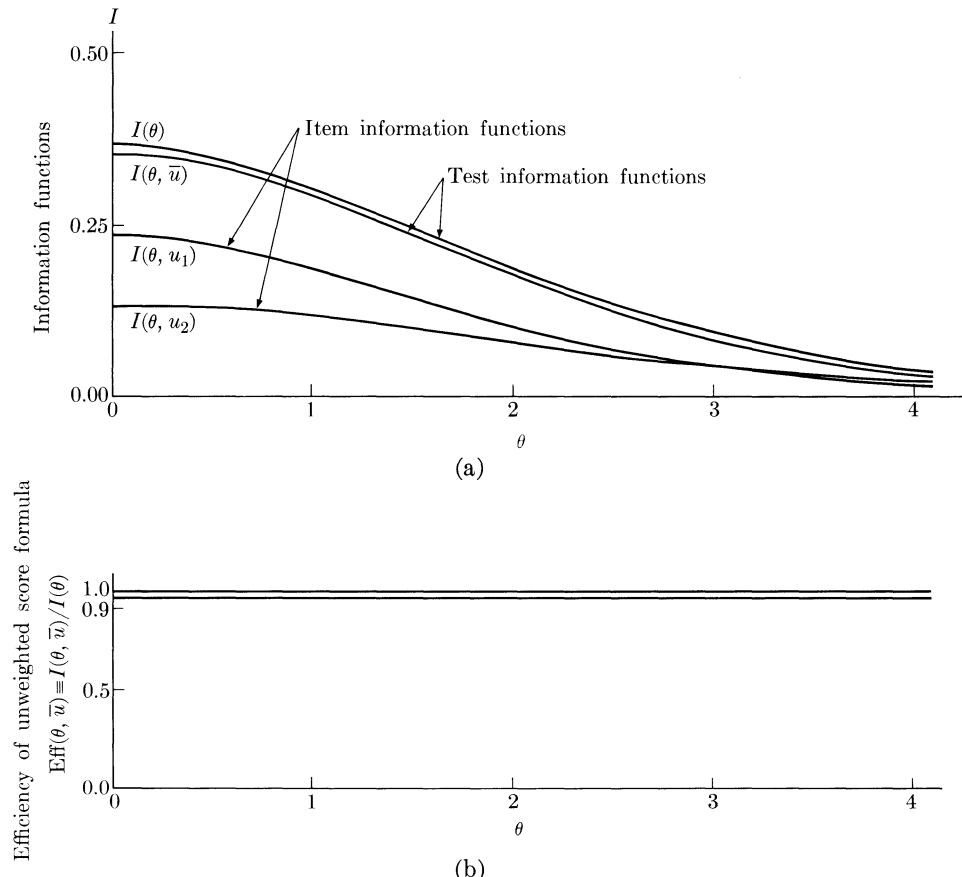


FIG. 20.8.3. Part (a) shows the two item information functions  $I(\theta, \mu_1)$  and  $I(\theta, \mu_2)$ . It also shows the two test information functions  $I(\theta)$  and  $I(\theta, \bar{u})$ , with unweighted and optimal weighted score formulas ( $x = 0.578 \mu_1 + 0.435 \mu_2$ ).

We observe that adding the second (type of) item and using  $\bar{u}$  causes an appreciable *reduction* in the values of the information function.

Figure 20.8.1b gives the efficiency

$$\text{Eff}(\theta, \bar{u}) = I(\theta, \bar{u})/I(\theta) \quad (20.8.7)$$

of  $\bar{u}$  in any logistic test model in which all the  $b_g$  are zero, and half the items have each of the  $a_g$ -values 0.98 and 0.204.

Figure 20.8.2 gives similar comparisons for items with  $a_1 = 0.6/\sqrt{0.64} = 0.75$  and  $a_2 = 0.3/\sqrt{0.91} \doteq 0.315$ ; and Fig. 20.8.3, similarly, for items with  $a_1 = 0.5/\sqrt{0.75} \doteq 0.578$  and  $a_2 = 0.4/\sqrt{0.84} \doteq 0.435$ . We see that in the case represented by Fig. 20.8.3, where relative variation between the  $a_g$  is small,

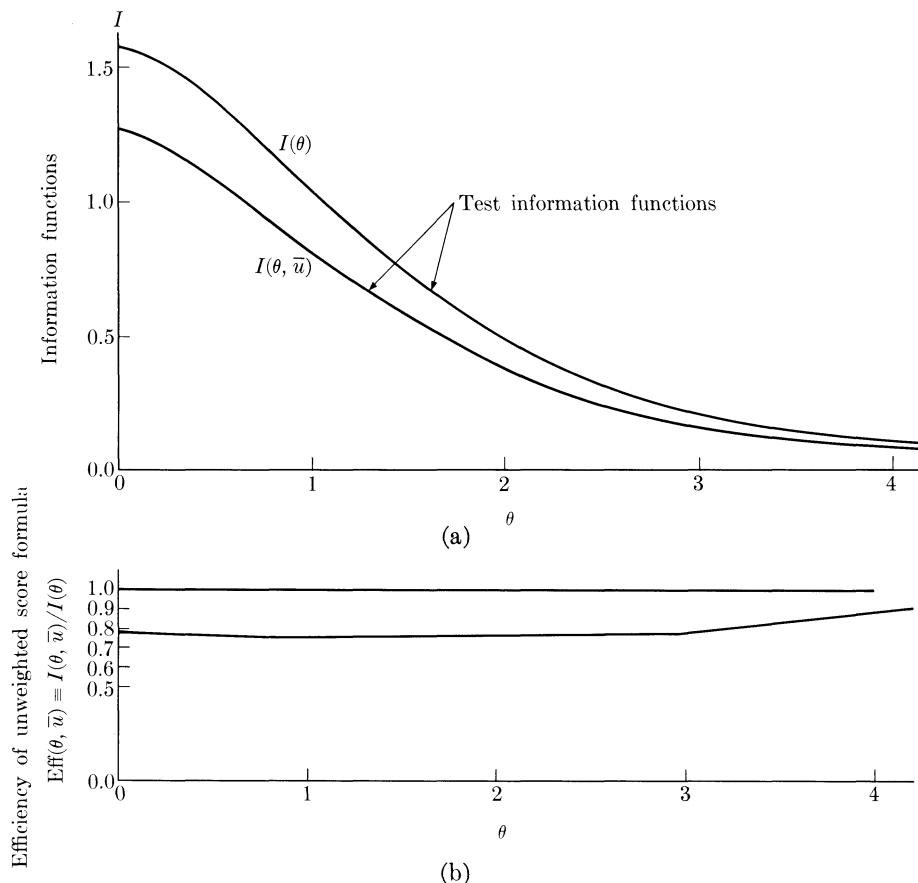


FIG. 20.8.4. Part (a) shows the two test item information functions  $I(\theta)$  and  $I(\theta, \bar{u})$ , with unweighted and optimal weighted score formulas ( $x = 0.98\mu_1 + 0.75\mu_2 + 0.578\mu_3 + 0.435\mu_4 + 0.315\mu_5 + 0.204\mu_6$ ).

the precision loss due to the use of unweighted scores is minor. In the intermediate case represented by Fig. 20.8.2, a major part of the information contributed by the second item is lost if the unweighted score is used. For the ability range ordinarily of interest, use of  $\bar{u}$  rather than  $x(v) = \sum_{g=1}^n a_g u_g$  is equivalent, in a test such as that represented in Fig. 20.8.1, to discarding about one-third of the information available; in a test such as that represented in Fig. 20.8.2, this is equivalent to discarding about one-sixth of the information available; and, in Fig. 20.8.3, to about 3%.

Figure 20.8.4 gives the information function of a hypothetical logistic test model of six items, with all the  $b_g$  equal to zero, and with the  $a_g$  respectively equal to the six values represented above, namely, 1, 0.75, 0.578, 0.435, 0.315,

and 0.204; it also gives the information function of the unweighted score formula  $\bar{u}$  in the same model, and the efficiency of  $\bar{u}$ , which is roughly 80% in the ability range usually of interest.

### References and Selected Readings

- BERKSON, J., Tables for the maximum likelihood estimate of the logistic function. *Biometrics*, 1957, **13**, 28-34.
- BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58-16*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January 1957.
- BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)
- BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17*. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)
- BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics*, 1958, **29**, 1284 (abstract). (c)
- BIRNBAUM, A., Statistical theory of tests of a mental ability. *Op. cit.*, 1285 (abstract). (d)
- BIRNBAUM, A., A unified theory of estimation, I. *Annals of Mathematical Statistics*, 1961, **32**, 112-135. (a)
- BIRNBAUM, A., The theory of statistical inference. New York: Institute of Mathematical Sciences, New York University, 1961. (b) (Mimeographed)
- BIRNBAUM, A., Statistical theory for logistic mental test models with a prior distribution of ability. *Research Bulletin 67-12*. Princeton, N.J.: Educational Testing Service, 1967.
- CRAMÉR, H., *Mathematical methods of statistics*. Princeton, N.J.: Princeton University Press, 1946.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*, Vol. 2. London: Griffin, 1961.
- LAWLEY, D. N., On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 1943, **61-A**, 273-287.
- LOEVINGER, JANE, The attenuation paradox in test theory. *Psychological Bulletin*, 1954, **51**, 493-504.
- LORD, F. M., A theory of test scores. *Psychometric Monograph*, No. 7. Chicago: University of Chicago Press, 1950.
- LORD, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17**, 181-194.
- LORD, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18**, 57-76.

- SITGREAVES, ROSEDITH, A statistical formulation of the attenuation paradox in test theory. In H. Solomon (Ed.), *Studies in item analysis and prediction*. Stanford: Stanford University Press, 1961.
- SOLOMON, H., Probability and statistics in psychometric research: item analysis and classification techniques. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, 5, pp. 169-184. Berkeley: University of California Press, 1956.
- TUCKER, L. R, Maximum validity of a test with equivalent items. *Psychometrika*, 1946, 11, 1-14.
- WALD, A., Asymptotically shortest confidence intervals. *Annals of Mathematical Statistics*, 1942, 13, 127-137.



**Part 6**

**STRONG TRUE-SCORE THEORY**

# POISSON PROCESS MODELS

## 21.1 Introduction

In this chapter, we shall study two Poisson process models that are of use in analyzing results from certain specialized kinds of tests. The presentation here is a condensation (with some refinements) of some of the work reported by Georg Rasch in his monograph *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960, now out of print). A third model proposed by Rasch has been mentioned previously as a special case of Birnbaum's logistic response model (see Section 17.3), in which all items are assumed to have equal discriminating power. The reader might consult Rasch (1966a) and Rasch (1966b); both these sources present this model with concise descriptions of its special properties.

Rasch proposed the first of his models as a tool for evaluation of one aspect of reading ability on the basis of the number of an examinee's misreadings in an oral reading test. In the testing procedure for which he proposed this model, the examinee is presented a text which he is required to read aloud, and a record is made of the number of words of the total text that he misreads. It is assumed that a person's probability of misreading any word is a small constant depending on the person but not on the particular word, and that these probabilities are independent over words for a given person. From these assumptions, Rasch derived a Poisson distribution for the number of misreadings as a model for this test.

Rasch also proposed a second model for evaluating a second aspect of reading ability on the basis of either the number of words an examinee reads during some fixed time period, or on his reading time, if the examinee finishes reading the text before the end of the fixed time period. In both these cases, the analysis may be based on the average number of words read in the actual time period used. Adopting assumptions similar to those of the first model, Rasch derived Poisson and gamma distributions for the two cases of this model. Both of the Rasch models are referred to as Poisson process models because of the basic similarity of the assumptions underlying them.

A major characteristic of the Rasch models is that each is a two-parameter model with one parameter identified with the ability of the person and the second parameter identified with the difficulty of the measurement. Further-

more, the ability and difficulty parameters are “separable” in the Rasch models, so that they may be estimated independently in a manner quite analogous to the way in which the parameters are estimated in a two-way factorial analysis of variance when it is assumed that there is no interaction. Finally, certain parameter-free distributions are available that make an investigation of the accuracy of the model possible.

In his writings, Rasch has emphasized that he is not concerned with distributions over people, but rather with estimation of ability for each person. Thus standard measures such as reliability and validity coefficients, which have meaning only in terms of distributions of scores over people, are of no interest to him.

Rasch (1960, p. 115) has stated that when these Poisson process models apply, they provide “a principle of measurement on a ratio scale of both difficulty and ability parameters, the conceptual status of which is comparable to that of measuring mass and force”. We shall discuss this “principle of measurement” and a difficulty of its application. It is not our intention to carry our presentation of these models to the point where the reader can use it as a guide to applications. Space limitations prevent our doing any more than exhibiting the major features of Rasch’s considerable body of work. Our purpose here is to show how it is possible, at least for simple tests, to derive a theoretical model for test scores and to evaluate logically the validity of this model in terms of the assumptions required for this derivation. We feel that this kind of “logical validity analysis”, suggested in Chapters 1 and 2, is essential to the justification of strong true-score models. In Chapters 22 and 23 we shall study other such models, again with some emphasis on the assumptions underlying them.

## 21.2 Generating Functions and the Poisson Distribution

We shall base the mathematical development of the model for misreadings on the theory of probability generating functions. This will facilitate the general development, and will also allow us to relax the assumptions underlying the model from those given by Rasch. Suppose we have a random variable taking nonnegative integral values  $i$  with probabilities  $p_i$ ,  $i = 0, 1, 2, \dots$ . Consider the probability generating function

$$G(t) = \sum_{i=0}^{\infty} p_i t^i. \quad (21.2.1)$$

Clearly

$$p_i = (i!)^{-1} \frac{\partial G^{(i)}(t)}{\partial t^{(i)}} \Big|_{t=0}, \quad i = 0, 1, 2, \dots; \quad (21.2.2)$$

that is, the  $i$ th derivative of the probability generating function evaluated at  $t = 0$ , divided by  $i!$ , is just  $p_i$ , the probability of the integer  $i$ . As an example,

consider the Poisson density function

$$h(i | \lambda) \equiv e^{-\lambda} \lambda^i / i!, \quad i = 0, 1, \dots, \quad (21.2.3)$$

with intensity parameter  $0 < \lambda < \infty$ . This density has the probability generating function

$$G(t) = \sum_{i=0}^{\infty} \frac{e^{-\lambda} \lambda^i}{i!} t^i = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda t)^i}{i!} = e^{-\lambda} e^{\lambda t} = e^{-\lambda(1-t)}. \quad (21.2.4)$$

Now, using superscripts to denote successive derivatives,

$$\begin{aligned} G^{(1)}(t) &= e^{-\lambda} \lambda e^{\lambda t} = \lambda e^{-\lambda(1-t)}, \quad \text{with } G^{(1)}(0)/1! = \lambda e^{-\lambda}, \\ G^{(2)}(t) &= \lambda^2 e^{-\lambda(1-t)}, \quad \text{with } G^{(2)}(0)/2! = \lambda^2 e^{-\lambda}/2!; \end{aligned}$$

and in general, the  $i$ th derivative is

$$G^{(i)}(t) = \lambda^i e^{-\lambda(1-t)}, \quad \text{with } G^{(i)}(0)/i! = \lambda^i e^{-\lambda}/i!.$$

Now let  $X$  and  $Y$  be independent, nonnegative, integer-valued random variables with probability distributions

$$\text{Prob } \{X = j\} = a_j \quad \text{and} \quad \text{Prob } \{Y = k\} = b_k.$$

Then the event  $(X = j, Y = k)$  has probability  $a_j b_k$ . The sum  $S = X + Y$  is a new random variable, and the event  $S = r$  is the union of mutually exclusive events

$$\begin{aligned} (X = 0, Y = r), \quad (X = 1, Y = r - 1), \\ (X = 2, Y = r - 2), \quad \dots, \quad (X = r, Y = 0). \end{aligned}$$

Therefore the distribution  $\text{Prob } \{S = r\}$  is given by

$$c_r = a_0 b_r + a_1 b_{r-1} + a_2 b_{r-2} + \dots + a_r b_0,$$

and we have

**Theorem 21.2.1.** If  $X$  and  $Y$  are independent, nonnegative, integer-valued random variables with probability generating functions

$$A(t) = \sum a_j t^j \quad \text{and} \quad B(t) = \sum b_k t^k,$$

where  $a_j = \text{Prob } \{X = j\}$  and  $b_k = \text{Prob } \{Y = k\}$ , then the probability generating function

$$C(t) = \sum c_m t^m$$

of  $S = X + Y$  is the product of the probability generating functions of  $X$  and  $Y$ ; that is,

$$C(t) = A(t)B(t).$$

*Proof.* Essentially, the theorem is proved, as indicated, by expanding  $A(t)$  and  $B(t)$  in series, obtaining the product of these series, and equating this, term by term, with the terms of  $C(t)$ .  $\square$

As an application of this theorem, we state

**Theorem 21.2.2.** If  $X_n$  are independent Poisson random variables with parameters  $\lambda_n$ , then  $S = \sum X_n$  has the Poisson distribution, with parameter  $\sum \lambda_n$ .

*Proof.* The probability generating function of each  $X_n$  is

$$G_n(t) = e^{-\lambda_n(1-t)},$$

and the generating function of  $S$  is

$$G_S(t) = \prod_n e^{-\lambda_n(1-t)} = e^{-\sum \lambda_n(1-t)}, \quad (21.2.5)$$

which is the probability generating function of a Poisson random variable with parameter  $\sum \lambda_n$ . The result then follows because of the one-to-one correspondence that exists between probability distributions and generating functions.  $\square$

A standard property of Poisson random variables that is central to the Rasch models concerns the conditional distribution of independent Poisson variables, given that their sum is a fixed quantity. We shall first develop this theory for the two-variable case, and then simply state the general result. We begin with a very standard result.

**Lemma 21.2.3.** Let  $Y_1$  and  $Y_2$  be independent Poisson random variables with parameters  $\lambda_1$  and  $\lambda_2$ , and let  $X = Y_1 + Y_2$ . Then the conditional distribution of  $Y_1$ , given  $X = x$ , is binomial, with parameters

$$x \quad \text{and} \quad \lambda_1/(\lambda_1 + \lambda_2);$$

that is,

$$P(y_1 | x) = \binom{x}{y_1} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{y_1} \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{x-y_1}. \quad (21.2.6)$$

*Proof.* We have

$$P(y_1 | x)P(x) = P(y_1, x).$$

But there exists a one-to-one correspondence between the points  $(y_1, x)$  and  $(y_1, y_2)$ , since  $x = y_1 + y_2$ . Likewise there is a one-to-one correspondence between the associated probabilities. Hence

$$P(y_1 | x) = \frac{P(y_1, y_2)}{P(x)}.$$

Now

$$P(y_1, y_2) = \frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!} \frac{e^{-\lambda_2} \lambda_2^{y_2}}{y_2!}$$

and

$$P(x) = \frac{e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^x}{x!}.$$

The result follows on evaluating  $P(y_1, y_2)/P(x)$ .  $\square$

We define the usual multinomial coefficient

$$\binom{x}{y_1 \ y_2 \ \dots \ y_n} \equiv \frac{x!}{y_1! \ y_2! \ \dots \ y_n!}.$$

Then an equivalent form for (21.2.6) is

$$P(y_1 | x) = \binom{x}{y_1 \ y_2} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{y_1} \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{y_2}, \quad (21.2.7)$$

where  $y_2 = x - y_1$ .

We now state the general result:

**Theorem 21.2.4.** Let  $Y_1, Y_2, \dots, Y_n$  be independent Poisson random variables with parameters  $\lambda_1, \lambda_2, \dots, \lambda_n$ , and let

$$X = Y_1 + Y_2 + \dots + Y_n.$$

Then the conditional distribution of  $Y_1, Y_2, \dots, Y_n$ , given  $X = x$ , is multinomial, with parameters

$$x \quad \text{and} \quad \frac{\lambda_1}{\sum \lambda_i}, \frac{\lambda_2}{\sum \lambda_i}, \dots, \frac{\lambda_n}{\sum \lambda_i}.$$

In other words,

$$P(y_1, y_2, \dots, y_n | x) = \binom{x}{y_1 \ y_2 \ \dots \ y_n} \left( \frac{\lambda_1}{\sum \lambda_k} \right)^{y_1} \left( \frac{\lambda_2}{\sum \lambda_k} \right)^{y_2} \dots \left( \frac{\lambda_n}{\sum \lambda_k} \right)^{y_n}, \quad (21.2.8)$$

where  $\sum \lambda_k = \sum_{k=1}^n \lambda_k$ . Actually

$$y_n = x - \sum_{i=1}^{n-1} y_i,$$

so that (21.2.8) is equivalent to

$$\begin{aligned} P(y_1, y_2, \dots, y_{n-1} | x) \\ = \binom{x}{y_1 \ y_2 \ \dots \ y_{n-1}} \left( \frac{\lambda_1}{\sum \lambda_i} \right)^{y_1} \left( \frac{\lambda_2}{\sum \lambda_i} \right)^{y_2} \dots \left( 1 - \frac{\sum_{i=1}^{n-1} \lambda_i}{\sum \lambda_i} \right)^{x - \sum_{i=1}^{n-1} y_i}. \end{aligned} \quad (21.2.9)$$

### 21.3 Derivation of the Model for Misreadings

#### as a Poisson Limit of Bernoulli Trials with Variable Probabilities

Following Rasch himself, we might base our derivation of the Rasch model for misreadings on the standard derivation of the Poisson distribution as the limit of a sum of independent and identically distributed Bernoulli random variables. However, because of the kinds of data that the model is suggested to describe, it seems useful to give the derivation for the equally standard but more inclusive case in which the Bernoulli parameters are variable.

Let  $Y_g$  be the random variable corresponding to an error ( $Y_g = 1$ ) or non-error ( $Y_g = 0$ ) on trial  $g$ . Let  $X$  be the total number of errors on  $n$  trials. Let

$$\text{Prob } \{Y_g = 1\} = p_g \quad \text{and} \quad \text{Prob } \{Y_g = 0\} = q_g = 1 - p_g.$$

The probability generating function of  $Y_g$  is  $(q_g + p_g t)$ , and the probability generating function of  $X$  is

$$P(t) = \prod_{g=1}^n (q_g + p_g t).$$

The logarithm of the probability generating function is

$$\log P(t) = \sum_{g=1}^n \log [1 - p_g(1 - t)].$$

Now, for very small  $x$ ,  $\log(1 - x) \doteq -x - \theta x$ , where  $\theta \rightarrow 0$  as  $x \rightarrow 0$ . Hence for small  $p_g$ ,

$$\sum_{g=1}^n \log [1 - p_g(1 - t)] = -(1 - t) \sum_{g=1}^n (p_g + \theta_g p_g).$$

Now let  $n \rightarrow \infty$  in such a way that the largest  $p_g$  tends to zero, but  $\sum_{g=1}^n p_g$  tends to  $\lambda$ ; then

$$\lim_{n \rightarrow \infty} \log P(t) = -\lambda(1 - t).$$

Thus

$$\lim_{n \rightarrow \infty} P(t) = e^{-\lambda + \lambda t},$$

which is the generating function of the Poisson distribution with parameter  $\lambda$ . Hence we have

**Theorem 21.3.1.** Let  $Y_g$ ,  $g = 1, 2, \dots, n$ , be independent Bernoulli random variables with probabilities (of error)  $P_g$  and suppose that as  $n \rightarrow \infty$ , the largest  $p_g \rightarrow 0$  and  $\sum_{g=1}^n p_g \rightarrow \lambda$ . Then the sum

$$X = \sum_{g=1}^n Y_g$$

approaches a Poisson random variable with parameter  $\lambda$  as  $n \rightarrow \infty$ .

Now suppose that the reading of each word of a text constitutes an independent Bernoulli trial, with probability of error on that word being some very small fraction that depends on the difficulty of the word. Because of the preceding theorem, we can expect the distribution of the total number of misreadings in the text, for large  $n$ , to be approximately a Poisson distribution, with intensity parameter  $\lambda$  equal to the sum of the Bernoulli parameters.

## 21.4 Statement of the Poisson Model for Misreadings

Suppose we are given the number of misreadings  $x_{ga}$  of  $n$  examinees on  $N$  texts. We wish to estimate the ability of each examinee and the difficulty of each text. Let  $\theta_{ga}$  be the probability of an error by examinee  $a$  on a fixed word in  $g$ . We assume that for each  $g$ ,  $\theta_{ga}$  takes on all values  $0 \leq \theta_{ga} \leq 1$  in the population of persons  $a$ . We assume that we can express  $\theta_{ga}$  as the ratio

$$\theta_{ga} = \delta_g / \xi_a \quad (21.4.1)$$

of a test (or item) difficulty parameter  $\delta_g$  and an ability parameter  $\xi_a$ . These parameters are not uniquely determined, for we could choose  $\delta'_g = 100\delta_g$  and  $\xi'_a = 100\xi_a$  and the ratio  $\theta_{ga}$  would remain unchanged. The scale of  $\theta_{ga}$  is determined by the relative frequency interpretation which it is most convenient to preserve for it. Note that if we choose a scale for one parameter, the scale of the other is thereby determined.

To fix the scale of the difficulty and ability parameters, Rasch selects the most difficult test in the *given* set, denotes it as test zero, refers to it as the reference test, and assigns it difficulty parameter  $\delta_0 = 1$ . Then for any person,

$$\xi_a = \theta_{0a}^{-1}; \quad (21.4.2)$$

the ability of a person is the reciprocal of his probability of a misreading in the reference test. Note that  $\xi_a$  takes values from unity to infinity, since  $\theta_{0a}$  takes values from zero to unity.

Rasch then selects as the reference person, some person with ability parameter  $\xi_a = 1$ , the lowest possible value. Such a person is certain to misread a random word on the reference test. He denotes this person by  $a = 0$ . Then from (21.4.1), for any test,

$$\theta_{g0} = \delta_g; \quad (21.4.3)$$

*The difficulty of a test  $g$  is the probability of a misreading by the standard person on the test.* Clearly  $0 \leq \delta_g \leq 1$ . Now the difficulty of the reference test has been specified as  $\delta_0 = 1$ , that is, the reference test is the most difficult test. Hence the claimed ratio scaling requires the determination (or specification) of the test of maximum difficulty. If, in some finite collection of tests, one test is most difficult, then it is always possible that a new, most difficult test might sometime be added to the collection. If this happens, then this new test must be

specified to be the new reference test and all values  $\xi_a$  and  $\delta_g$  must be recomputed with reference to the new standard. From this point of view, the claimed ratio scaling seems less than completely determined in practice.

Assuming that the values  $\theta_{ga}$  are constant within a test, we may write

$$\lambda_{ga} = N_g \theta_{ga} = N_g \delta_g / \xi_a, \quad (21.4.4)$$

assuming that the examinee reads all  $N_g$  words in test  $g$ . We may call the factor  $N_g \delta_g$  the *impediment* of the text, and denote it by

$$\tau_g \equiv N_g \delta_g. \quad (21.4.5)$$

This allows us to write the expected number of misreadings of an examinee  $a$  on test  $g$  as

$$\lambda_{ga} = \tau_g \xi_a; \quad (21.4.6)$$

for a test of length  $N_g$ , the parameter  $\lambda_{ga}$  of the Poisson law is the ratio of the impediment of the test to the ability of the examinee.

If independent measurements  $g$  and  $h$ , having Poisson distributions with intensity parameters  $\lambda_{ga} = \tau_g / \xi_a$  and  $\lambda_{ha} = \tau_h / \xi_a$ , are administered to an examinee  $a$ , then the sum of the observed scores on the two tests has a Poisson distribution with intensity parameter

$$\lambda_{+a} = \lambda_{ga} + \lambda_{ha} = (\tau_g + \tau_h) / \xi_a. \quad (21.4.7)$$

This result may be referred to as *the additivity of impediments*. Thus the sum of two independent Poisson measurements is equivalent to one measurement, the impediment of which is the sum of the impediments of the two tests. The generalization to more than two measurements is immediate, yielding

**Theorem 21.4.1.** Let  $Y_{ga}$ ,  $g = 1, 2, \dots, n$ , be independent Poisson measurements with intensity parameters  $\lambda_{ga} = \tau_g / \xi_a$  for a specified person  $a$ . Then the total score  $X = \sum_{g=1}^n Y_{ga}$  has a distribution equivalent to a measurement having a Poisson distribution with parameter  $\sum \tau_g / \xi_a$ .

## 21.5 The Separation of Parameters in the Model for Misreadings

The special feature of the Rasch models is that they permit independent estimation of the ability and difficulty parameters. In this section, we shall illustrate the separation of parameters by which this is accomplished for the model for misreadings.

For experimentally independent Poisson measurements  $g$  and  $h$ , the joint distribution of measurements on a specified person  $a$  is given by

$$h(y_{ga}, y_{ha} | \tau_g, \tau_h, \xi_a) = \frac{e^{-(\tau_g + \tau_h)/\xi_a} \tau_g^{y_{ga}} \tau_h^{y_{ha}}}{y_{ga}! y_{ha}! \xi_a^{y_{+a}}}, \quad (21.5.1)$$

where  $y_{+a} = y_{ga} + y_{ha}$ . Note that by the factorization theorem (see Section 18.2),  $Y_{+a}$  is sufficient for  $\xi_a$ . We also know that the distribution of  $Y_{+a}$  is Poisson with parameters  $\tau_+ \equiv \tau_g + \tau_h$  and  $\xi_a$ :

$$h(y_{+a} | \tau_g, \tau_h, \xi_a) = \frac{e^{-\tau_+/\xi_a} (\tau_+/\xi_a)^{y_{+a}}}{y_{+a}!}. \quad (21.5.2)$$

Now we apply Lemma 21.2.3, with

$$\lambda_1 = \lambda_{ga} = \tau_g/\xi_a, \quad \lambda_2 = \lambda_{ha} = \tau_h/\xi_a. \quad (21.5.3)$$

Then

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\tau_g/\xi_a}{\tau_g/\xi_a + \tau_h/\xi_a} = \frac{\tau_g}{\tau_g + \tau_h} = \frac{\tau_g}{\tau_+} \quad \text{and} \quad \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{\tau_h}{\tau_+}. \quad (21.5.4)$$

Hence the condition density of  $Y_{ga}$ , given  $y_{+a}$ , is

$$h(y_{ga} | y_{+a}) = \binom{y_{+a}}{y_{ga}} \rho_g^{y_{ga}} \rho_h^{y_{ha}}, \quad (21.5.5)$$

where  $\rho_g = \tau_g/\tau_+$  and  $\rho_h = \tau_h/\tau_+$  are the *relative impediments*. We note that this conditional distribution does not depend on  $\xi_a$ .

An estimate of the relative impediment  $\rho_g$  can be obtained from each of the groups of examinees with different values  $y_{+a}$ . It is also possible to obtain a pooled estimate of each of these relative impediments. This pooling may be effected by noting that the sum  $y_{g+}$  over  $a$ , given the sum  $y_{++}$  of  $y_{ga}$  over  $g$  and  $a$ , is binomially distributed with parameters  $y_{g+}$  and  $\rho_g$ :

$$h(y_{g+} | y_{++}) = \binom{y_{++}}{y_{g+}} \rho_g^{y_{g+}} \rho_h^{y_{h+}}. \quad (21.5.6)$$

We state this result more generally for  $n$  Poisson sums as

**Theorem 21.5.1.** The joint distribution of the sums

$$y_{g+}, y_{h+}, y_{i+}, \dots, y_{n+}, \quad \text{given } y_{++},$$

is the multinomial

$$h(y_{g+}, y_{h+}, \dots, y_{n+} | y_{++}) = \binom{y_{++}}{y_{g+} \ y_{h+} \ \dots \ y_{n+}} \rho_g^{y_{g+}} \rho_h^{y_{h+}} \cdots \rho_n^{y_{n+}}. \quad (21.5.7)$$

The estimation of the ability parameters independently of the difficulty parameters is also easily accomplished. Let  $\alpha_a = \xi_a^{-1}$ , and let  $\mu_a = \alpha_a/\alpha_+$ , where

$$\alpha_+ = \sum_{a=1}^n \alpha_a.$$

Then it is readily determined that

$$h(y_{+a}, y_{+b}, \dots, y_{+N} | y_{++}) = \binom{y_{++}}{y_{+a} \ y_{+b} \ \dots \ y_{+N}} \mu_a^{y_{+a}} \mu_b^{y_{+b}} \cdots \mu_N^{y_{+N}}. \quad (21.5.8)$$

Standard estimation techniques for estimating multinomial parameters may then be used to estimate the relative inability parameters  $\mu_a$ . Thus item difficulty and examinee ability parameters may be estimated independently of each other.

## 21.6 A Parameter-Free Distribution for Testing the Fit of the Model for Misreadings

Finally, let  $\mathbf{X} = \|x_{ga}\|$  be the matrix of random variables  $X_{ga}$ , let  $\mathbf{x}_{+a}$  be the vector of row totals  $x_{+a} = \sum_g x_{ga}$ , and let  $\mathbf{x}_{g+}$  be the vector of column totals  $x_{g+} = \sum_a x_{ga}$ . Then Rasch shows that

$$\text{Prob } \{\mathbf{X} | \mathbf{x}_{+a}, \mathbf{x}_{g+}\} = \frac{x_{++}!}{\binom{x_{++}}{\mathbf{x}_{+a}} \binom{x_{++}}{\mathbf{x}_{g+}} \prod_g \prod_a x_{ga}!} = \frac{\prod_g x_{g+}! \prod_a x_{+a}!}{x_{++}! \prod_g \prod_a x_{ga}!}, \quad (21.6.1)$$

where

$$\binom{x_{++}}{\mathbf{x}_{+a}} \equiv \binom{x_{++}}{x_{+1} \ x_{+2} \ \cdots \ x_{+N}} \quad \text{and} \quad \binom{x_{++}}{\mathbf{x}_{g+}} \equiv \binom{x_{++}}{x_{1+} \ x_{2+} \ \cdots \ x_{n+}}.$$

Rasch briefly discusses methods, based on this distribution, for testing the fit of the model: From the given data, and independently of the true parameter values, it is possible to examine the closeness of fit of the data to the model.

## 21.7 A Stochastic Process Model for Oral Reading Speed

Rasch's second model is concerned with the speed of oral reading. The pupil is given a text to be read aloud, and a record is made of the amount of time  $t$  that is required to read  $n$  words. Rasch assumes that the probability that an examinee will complete the reading of a word in any small time interval depends only on the length  $\Delta t$  of that interval, and not, for example, on the particular word nor on the length of time that has passed since the reading of the last word was completed. He further assumes that as  $\Delta t \rightarrow 0$ , the probability of more than one completion occurring in the interval is negligible in comparison with the probability of zero or one occurrences. Under these assumptions, the Poisson distribution may be derived by direct reference to the underlying continuous process (see Feller, 1957, pp. 400ff.). Though this derivation is more informative, Rasch's derivation is simpler mathematically. Rasch assumes a binomial model and writes the probability of  $y$  readings being completed in

$N$  time periods as

$$\text{Prob } \{y | N\} = \binom{N}{y} p^y (1 - p)^{N-y}, \quad (21.7.1)$$

where  $p$  is the probability of a reading being completed in any single time period.

Then Theorem (21.3.1) may be applied to deduce that the binomial model may be approximated by the Poisson model

$$\text{Prob } \{y | N\} = \frac{(Np)^y}{y!} e^{-Np},$$

where  $Np$  is the expected number of readings completed in the period. In applying this theorem here, it should be noted that the values of  $p$  being considered are large in relation to the probabilities of error in the preceding model. Hence the oral reading model should not be expected to be as good as the model for misreadings.

To emphasize the continuous nature of the time variable, we write  $Np$  as  $\lambda t$  in the limit, where  $\lambda$  is the average number of words read per unit of time. Thus the Poisson distribution for the number of words  $y$  an examinee reads in the total time  $t$  is

$$\text{Prob } \{y | t\} = \frac{(\lambda t)^y}{y!} e^{-\lambda t}. \quad (21.7.2)$$

In his derivation of this distribution Rasch assumes that the text is homogeneous, so that the value  $p$  is a constant for all intervals. This assumption is actually unnecessary since, for application of Theorem 21.3.1, the value  $p$  can be taken to be the average probability of a reading completion over intervals. Hence in this model, as in the model for misreadings, the homogeneity assumption is unnecessary. Rather, the important assumption in each case is that the average  $p$  value be very small.

## 21.8 Uncompleted Texts

Since some examinees may complete the entire text, it is sometimes more convenient to permit all examinees to complete the entire text and to record and analyze the amount of time they take to do this. The probability model for reading rate in this measurement procedure can be obtained directly from (21.7.2). Note, as did Rasch (1960, p. 38), that “the event that the number of words  $a$  reads in a given time  $T$  exceeds a given number  $N$  is identical to the event that the time  $t$  used to read  $N$  words is less than  $T$ ”. Symbolically,

$$\text{Prob } \{a \geq N | t\} = \text{Prob } \{t \leq T | N\}. \quad (21.8.1)$$

The left-hand member of (21.8.1) is the sum of the probabilities that

$$a = N, \quad N + 1, \quad N + 2, \quad \dots$$

in (21.7.2). Therefore we have

$$\text{Prob } \{t \leq T | N\} = e^{-\lambda t} \left( \frac{(\lambda t)^N}{N!} + \frac{(\lambda t)^{N+1}}{(N+1)!} + \dots \right), \quad (21.8.2)$$

which is the cumulative distribution of the variable  $t$ . Note that if  $N = 0$ ,

$$\begin{aligned} \text{Prob } \{t \leq T | 0\} &= e^{-\lambda t} \left( 1 + \lambda t + \frac{(\lambda t)^2}{2!} + \dots \right) \\ &= e^{-\lambda t} e^{\lambda t} = 1. \end{aligned} \quad (21.8.3)$$

Also

$$\begin{aligned} \text{Prob } \{t \leq T | 1\} &= e^{-\lambda t} \left( \lambda t + \frac{(\lambda t)^2}{2!} + \dots \right) \\ &= e^{-\lambda t} (e^{\lambda t} - 1) = 1 - e^{-\lambda t}. \end{aligned} \quad (21.8.4)$$

By differentiating the cumulative distribution function (21.8.4), we obtain the probability density function

$$p\{t | 1\} = \lambda e^{-\lambda t}. \quad (21.8.5)$$

This is the probability density function of the reading time for the first word, and hence, because of the independence assumption, of any single word. The density (21.8.5) is a negative exponential density. The probability density function for the reading time for  $N$  words may be obtained by differentiating (21.8.2). This yields the density

$$p\{t | N\} = \lambda e^{-\lambda t} \frac{(\lambda t)^{N-1}}{(N-1)!}, \quad (21.8.6)$$

a gamma density, which reduces to (21.8.5) for  $N = 1$ .

As in the model for misreadings, Rasch characterizes the intensity parameter of the process as the ratio of two parameters, one associated with the difficulty of the text and the other associated with the ability of the examinee. A separation of parameters is again obtained, and thus it is possible to estimate each parameter independently of the other. The derivations, however, are more complex than for the model for misreadings. Rasch (1960) has given a more complete general discussion of the separability of parameters property.

Rasch has also reported some investigations of the empirical validity of these models. The evidence for empirical validity reported by Rasch (and others) is more positive for the model for misreadings than for the model for reading speed. The least favorable evidence is for the item analysis model mentioned in Chapter 19. These findings are not surprising if we view them in connection with the derivations of these models presented here. For the misreading model,  $p$  is small and the number of words is large, and hence the asymptotic theory would seem to be very appropriate for the test lengths encountered in applications. For the reading speed models  $p$  is substantially larger, and hence, as

is well known, a very much larger number of observations must be summed before the asymptotic theory applies.

With respect to discrete items, it appears that if the number of items is very large, then inferences about an examinee's ability based on his total test score will be very much the same whether Rasch's third model, which has no item discriminating power parameter, or Birnbaum's model, which does have an item discriminating power parameter, is used. Rasch's model in effect treats all items as having a discriminating power equal to the average discriminating power of the items. For the item analysis model, the  $p$  value is relatively large and the number of items of a test is typically far less than asymptotic theory would demand for the omission of the third parameter. This same comment applies, though with less force, to the model for reading speed. It is clear that more detailed theoretical and more extensive empirical examinations of each of the Rasch models would be very useful, since these models, when their conditions are satisfied, provide a mode of analysis of great simplicity and power.

### References and Selected Readings

- FELLER, W., *An introduction to probability theory and its applications*, Vol. I. New York: Wiley, 1957.
- RASCH, G., Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960.
- RASCH, G., On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Berkeley: University of California Press, 1961, 4, pp. 321-324.
- RASCH, G., An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in mathematical social science*. Chicago: Science Research Associates, 1966, pp. 89-107. (a)
- RASCH, G., An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, 19, 49-57. (b)

## CHAPTER 22

# MEASUREMENTS WITH INDEPENDENT ERRORS OF KNOWN DISTRIBUTIONAL FORM

### 22.1 Introduction

In classical test theory, it is assumed that errors of measurement are uncorrelated with each other and with all true scores. In test theory for  $\tau$ -equivalent measurements (Chapter 10), it is additionally assumed that the conditional expectation of any error of measurement on a given test is zero when other errors of measurement and the true score are fixed (Eq. 10.2.1). This strengthening of assumptions makes it possible to express higher-order moments of the latent variables in terms of observed-score moments.

The present chapter is based on

**Assumption 22.1.1.** *The errors of measurement on a test are distributed independently of true score on the same test.*

As in earlier chapters, the origin used for measuring true score is chosen so that the mean error of measurement is zero.

For tests composed of dichotomously scored items, the authors prefer the assumptions of Chapter 23, for the following reason: If the observed score on a test is a bounded variable, as it usually is, and if there are examinees with true scores near the “ceiling” or the “floor”, then it is implausible to assume that the errors of measurement are distributed independently of true score. If an examinee’s true score is near the ceiling, for example, his observed score cannot contain a large positive error, but it can contain a large negative error. If an examinee’s true score is near the floor, the opposite is true. In Section 10.8, we cited empirical results supporting such a conclusion. The assumption of independently distributed errors of measurement is useful (1) as a rough approximation, (2) when there is no floor or ceiling to the test, and (3) when there are no examinees near the floor or ceiling.

Under Assumption 22.1.1, the examinee’s true score is a location parameter, as pointed out by Novick (1966). If one has several measurements on a single examinee, many results are available for making statistical inferences (for example, Pitman, 1939, and Girshick and Savage, 1951).

In the present chapter we shall consider a different situation, in which there is only one measurement on each of a large number of examinees. The typical statistical inference problems here may be called *empirical Bayes* problems

(Robbins, 1956, 1963, 1964). For such problems, a wealth of conclusions can be drawn from Assumption 22.1.1. Many relevant results appear in the mathematical literature under the headings *convolutions* (Hirschman and Widder, 1955), or *mixtures of distributions* or *compound distributions* (Patil and Joshi, 1966; Patil, 1965; Teicher, 1960). As we shall frequently point out, although the mathematical problems relating to populations have been extensively and elegantly dealt with, the statistical inference problems that occur in practical work with samples drawn from these populations are largely unsolved. This is an area in which statisticians may produce significant developments in the near future.

Typical problems are (1) inferring the frequency distribution of true score (Sections 4 and 5), and (2) inferring the regression of true score on observed score (Sections 6, 8, and 9). If the true-score distribution can be estimated, then the entire bivariate distribution of true score and observed score can be estimated, including the regression of true score on observed score. This regression can of course be used to estimate the true score of a particular examinee from his observed score.

The present chapter uses a somewhat different type of mathematics than the preceding chapters, relying on the differential and integral calculus. The results obtained are instructive, but primarily of theoretical interest. Reading of this chapter may be omitted without loss of continuity. The reader interested only in routine practical applications should do so.

## 22.2 Moments of the True-Score Distribution\*

In this section, we shall derive a simple relationship that can be used to estimate true-score moments from observed-score moments. We have already discussed the practical uses of such estimated moments in Chapter 10.

A fundamental theorem relating to moment generating functions states that if all moments exist, the moment generating function (mgf) of a sum of independent random variables is equal to the product of their separate mgf's. (A more general theorem is available for characteristic functions; see, for example, Kendall and Stuart, 1958, Section 7.18.) Now by Assumption 22.1.1,

$$\Phi_X(\theta) = \Phi_T(\theta)\Phi_E(\theta), \quad (22.2.1)$$

where  $\Phi_X(\theta)$ ,  $\Phi_T(\theta)$ , and  $\Phi_E(\theta)$  are the mgf's of observed score, true score, and error of measurement, respectively.

The logarithm of the mgf, here denoted by  $\psi$ , is a cumulant† generating function. We have

$$\psi_X(\theta) = \psi_T(\theta) + \psi_E(\theta). \quad (22.2.2)$$

\* This section assumes some familiarity with cumulants and moment generating functions. Reading of the entire section may be omitted without loss of continuity.

† The  $r$ th cumulant is a polynomial function of the first  $r$  moments.

If the cumulant generating function of a variable can be expanded in a convergent power series in  $\theta$ , then the coefficient of  $\theta^r/r!$  will be the  $r$ th cumulant of the variable. Thus, from (22.2.2), if the expansions exist, then

$$\begin{aligned}\sum_{r=1}^{\infty} \kappa_X^{(r)} \frac{\theta^r}{r!} &= \sum_{r=1}^{\infty} \kappa_T^{(r)} \frac{\theta^r}{r!} + \sum_{r=1}^{\infty} \kappa_E^{(r)} \frac{\theta^r}{r!} \\ &= \sum_{r=1}^{\infty} (\kappa_T^{(r)} + \kappa_E^{(r)}) \frac{\theta^r}{r!},\end{aligned}\quad (22.2.3)$$

where  $\kappa^{(r)}$  denotes the cumulant of order  $r$ . Since the representation of a function by a convergent power series is unique, we have, finally,  $\kappa_X^{(r)} = \kappa_T^{(r)} + \kappa_E^{(r)}$ , or better,

$$\kappa_T^{(r)} = \kappa_X^{(r)} - \kappa_E^{(r)}, \quad r = 1, 2, \dots. \quad (22.2.4)$$

*If the expansions in (22.2.3) converge for some value of  $\theta$ , then the  $r$ th cumulant of the true-score distribution is equal to the  $r$ th cumulant of the observed-score distribution minus the  $r$ th cumulant of the distribution of the errors of measurement.*

If the distribution of the errors of measurement is completely known, then all the cumulants (and thus all the moments) of the true-score distribution can be obtained from the cumulants of the observed-score distribution by (22.2.4). For example, if the errors of measurement are distributed normally and independently of true score with mean zero and variance  $\sigma_E^2$ , then  $\kappa_E^{(2)} = \sigma_E^2$  and all other cumulants of the error distribution are zero. The result is

**Theorem 22.2.1.** *If the errors of measurement are distributed normally and independently of true score with mean zero and variance  $\sigma_E^2$ , and if the summation on the left-hand side of (22.2.3) exists for some value of  $\theta$ , then the cumulants of the true-score distribution are identical with those of the observed-score distribution except for the second cumulant, which is found from the usual formula  $\sigma_T^2 = \sigma_X^2 - \sigma_E^2$ .*

Practical application of this theorem requires the estimation of  $\sigma_E^2$  by the administration of parallel test forms or by some other method.

The corresponding theorem for many tests may be proved similarly:

**Theorem 22.2.2.** *Given the random variables  $X_g$ ,  $g = 1, 2, \dots, n$ , where  $X_g = T_g + E_g$  and the  $E_g$  are distributed normally and independently of each other and of the  $T_g$  with zero means and variances  $\sigma^2(E_g)$ : If the multivariate cumulant generating function for the  $X_g$  can be expanded in a convergent series in  $\theta$ , then each multivariate cumulant of the  $T_g$  is equal to the corresponding cumulant of the  $X_g$ , excepting only that*

$$\sigma^2(T_g) = \sigma^2(X_g) - \sigma^2(E_g), \quad g = 1, 2, \dots, n.$$

This theorem implies, for example, that  $\sigma(X_1, X_2) = \sigma(T_1, T_2)$ ; also further relationships for higher cumulants.

### 22.3 Basic Equations in Latent Trait Theory

An elementary theorem in mathematical statistics states that integration (suitably defined) of any bivariate distribution over one of the variables yields the (marginal) distribution of the other variable. For example, if  $\chi(x, \tau)$  is the bivariate distribution of  $X$  and  $T$ , and is a continuous function of  $T$ , then we obtain

$$\varphi(x) = \int_{-\infty}^{\infty} \chi(x, \tau) d\tau \quad \text{for all } x. \quad (22.3.1)$$

Since we shall not need to deal with discontinuous functions of  $\tau$ , we shall assume the continuity of functions of  $\tau$  throughout this chapter.

By another elementary theorem, any bivariate distribution can be written as the product of a marginal distribution and a conditional distribution. Thus, if  $g(\tau)$  is the (marginal) distribution of  $\tau$ , and  $h(x | \tau)$  is the conditional distribution of  $X$  for given  $\tau$ , then

$$\chi(x, \tau) \equiv g(\tau)h(x | \tau), \quad (22.3.2)$$

so that (22.3.1) can always be written

$$\varphi(x) = \int_{-\infty}^{\infty} g(\tau)h(x | \tau) d\tau \quad \text{for all } x. \quad (22.3.3)$$

This equation is basic to much of latent trait theory,  $X$  and  $T$  being construed as observed score and true score, respectively.

Since  $h(x | \tau)$  is the conditional distribution of  $x$ , a change of variable shows that  $h(\tau + e | \tau)$  is the conditional distribution of  $E = X - T$ , the error of measurement. By Assumption 22.1.1,  $h(\tau + e | \tau)$  is not a function of  $\tau$ ; hence we denote the (conditional or unconditional) frequency distribution of  $E$  by  $k(e)$ , writing

$$h(\tau + e | \tau) \equiv k(e). \quad (22.3.4)$$

A change of variable now converts (22.3.3) into

$$\varphi(x) = \int_{-\infty}^{\infty} g(x - e)k(e) de. \quad (22.3.5)$$

(In general mathematical usage, if three functions  $\varphi$ ,  $g$ , and  $k$  satisfy (22.3.5), it is said that  $\varphi$  is the *convolution* of  $g$  and  $k$ .)

Equation (22.3.5) is a starting point for solving various problems. We may state the most basic theoretical problem: For a given error distribution  $k$ , infer the true-score distribution  $g$  from the observed-score distribution  $\varphi$ .

If this inference can be made, then the bivariate distribution of observed score and true score can be reconstructed from (22.3.2). This bivariate distribution provides a complete description of the probabilistic relation of  $X$  to  $T$ , that is, a complete description of how a test functions as a measuring instrument.

An index can be constructed, for example, that describes the effectiveness of a test as a measuring instrument at different observed-score levels or at different true-score levels. If desired, the (curvilinear) regression of true score on observed score can be obtained from (22.3.2) and used to estimate any examinee's true score from his observed score. Other practical applications for the estimated true-score distribution are outlined in Chapter 23. One method for estimating the true-score distribution is given in the next section.

## 22.4 A Formula for the Distribution of True Scores in Terms of the Distribution of Observed Scores

Here we shall derive a result used in astronomy and other areas, a result originally obtained by Eddington (1913). [Readers uninterested in the derivation should skip directly to (22.4.8).] Other methods of dealing with this problem are given by Trumpler and Weaver (1953, Chapters 1.4 and 1.5).

Consider the problem of estimating  $g(\tau)$  from  $\varphi(x)$  in (22.3.3) or (22.3.5). Note that

$$\left. \frac{\partial^r}{\partial e^r} g(x - e) \right|_{e=0} = (-1)^r g^{(r)}(x), \quad (22.4.1)$$

where  $g^{(r)}(x)$  is the  $r$ th derivative of  $g(x)$ . We see that the expansion of  $g(x - e)$  in powers of  $e$ , if it exists, is

$$g(x - e) = g(x) - eg'(x) + \frac{e^2}{2!} g''(x) - \frac{e^3}{3!} g'''(x) + \dots \quad (22.4.2)$$

If (22.4.2) exists for all values of  $e$ , and if the necessary moments of the distribution of  $e$  are finite, then (22.4.2) can be multiplied by  $k(e)$  and integrated term-by-term. Together with (22.3.5), this means that

$$\varphi(x) = g(x) \int_{-\infty}^{\infty} k(e) de - g'(x) \int_{-\infty}^{\infty} ek(e) de + \frac{1}{2} g''(x) \int_{-\infty}^{\infty} e^2 k(e) de - \dots,$$

or (since the second integral is zero) that

$$\varphi(x) = g(x) + \frac{1}{2}\mu_2 g''(x) - \frac{1}{6}\mu_3 g'''(x) + \dots, \quad (22.4.3)$$

where  $\mu_r$  is the  $r$ th central moment of the distribution of  $e$ .

Before proceeding, it is worth noting that (22.4.3) throws some light on the effects of errors of measurement.

- 1) In any range of  $T$  where the true-score distribution is nearly a straight line, the observed-score and true-score distributions will, by (22.4.3), be nearly identical.
- 2) Since the second derivative  $g''(\tau)$  is negative at the mode(s) and positive at the antimodes (if any) of  $g(\tau)$ , Eq. (22.4.3) shows that the errors of measurement tend to round off the peaks of the true-score distribution and to fill in the valleys.

Differentiate (22.4.3)  $r$  times, obtaining

$$\varphi^{(r)}(x) = g^{(r)}(x) + \frac{1}{2}\mu_2 g^{(r+2)}(x) - \frac{1}{6}\mu_3 g^{(r+3)}(x) + \dots \quad (22.4.4)$$

Suppose now that (22.4.3) can be inverted to express  $g(x)$  in terms of  $\varphi(x)$  and its derivatives:

$$g(x) = \varphi(x) + A_1\varphi'(x) + A_2\varphi''(x) + \dots, \quad (22.4.5)$$

where  $\varphi'$ ,  $\varphi''$ ,  $\dots$ ,  $\varphi^{(r)}$  are derivatives of  $\varphi(x)$ , and  $A_1, A_2, \dots, A_r$  are coefficients to be determined. Substituting (22.4.3) and (22.4.4) into (22.4.5), we have

$$\begin{aligned} g(x) &= g(x) &+ \frac{1}{2}\mu_2 g''(x) - \frac{1}{6}\mu_3 g'''(x) &+ \frac{1}{24}\mu_4 g^{iv}(x) - \dots \\ &+ A_1g'(x) &+ \frac{1}{2}A_1\mu_2 g'''(x) - \frac{1}{6}A_1\mu_3 g^{iv}(x) + \dots \\ &+ A_2g''(x) &+ \frac{1}{2}A_2\mu_2 g^{iv}(x) - \dots \\ &+ A_3g'''(x) &+ \dots \\ &+ \dots &+ A_4g^{iv}(x) + \dots \end{aligned} \quad (22.4.6)$$

Now  $A_1, A_2, \dots$  can be determined so as to make this last equation an identity:

$$A_1 = 0, \quad A_2 = -\frac{1}{2}\mu_2, \quad A_3 = \frac{1}{6}\mu_3, \quad A_4 = \frac{1}{4}\mu_2^2 - \frac{1}{24}\mu_4, \quad \dots \quad (22.4.7)$$

Substituting (22.4.7) into (22.4.5) gives, finally, the result originally obtained by Eddington (1913):

$$g(x) = \varphi(x) - \frac{1}{2}\mu_2 \varphi''(x) + \frac{1}{6}\mu_3 \varphi'''(x) + (\frac{1}{4}\mu_2^2 - \frac{1}{24}\mu_4)\varphi^{iv}(x) + \dots \quad (22.4.8)$$

the  $\mu$  being the central moments of the error distribution.

Kurth (1965) gives necessary and sufficient conditions for the validity of (22.4.8). Sufficient conditions for the convergence of (22.4.8) to a solution of the integral equation (22.3.5) are that the derivatives of  $\varphi(x)$  of all orders exist and are suitably bounded, and that the characteristic function of  $k(e)$  has a reciprocal whose power series expansion in  $\theta$  converges for all values of  $\theta$ .

In the special case where the errors of measurement are normally distributed with variance  $\sigma_E^2$ , (22.4.8) simplifies to

$$\begin{aligned} g(x) &= \varphi(x) - \frac{1}{2}\sigma_E^2 \varphi''(x) + \frac{1}{8}\sigma_E^4 \varphi^{iv}(x) - \frac{1}{48}\sigma_E^6 \varphi^{vi}(x) + \dots \\ &= \sum_{k=0}^{\infty} \frac{(-\sigma_E^2)^k}{2^k k!} \varphi^{(2k)}(x). \end{aligned} \quad (22.4.9)$$

Pollard (1953) obtained a similar equation, using the method of integral transforms.

## 22.5 Statistical Inference Problems

Concerning (22.4.8), Trumpler and Weaver (1953, Section 1.43) stated:

Without going into a detailed discussion of the conditions for convergence, we may say that this series gives in general a good approximation with a moderate number of terms:

- i) when the values of the kernel [ $k(e)$ ] are appreciable only within a limited range on either side of the mean;
- ii) when the functions [ $g(x)$  and  $\varphi(x)$ ] have a ‘smooth run’ without high curvatures within any interval of  $x$  which is of the same extent as the ‘spread’ [range] of the kernel.\*

Equations (22.4.8) and (22.4.9) are usually modified for practical use by dropping all but the first two terms of the series, replacing the population distribution  $\varphi(x)$  by the frequency distribution of the (grouped) sample observations, and, finally, by approximating the required derivatives by the corresponding difference quotients (Scarborough, 1955, Section 101) of the sample frequency distribution.

Now it is clear that for some true-score distributions—for example, the Cauchy distribution—the power-series expansion of (22.4.2) does not converge for all  $x$  and  $e$ . Thus formulas (22.4.8) and (22.4.9) would not necessarily yield the correct true-score distribution even if  $\varphi(x)$  were actually known. It follows that the consistency of any sample estimator derived from (22.4.8) or (22.4.9) is open to question.

To the writers’ knowledge, very little work has been done on the statistical properties of sample estimators derived from (22.4.8) or (22.4.9), except for the work reported by Gaffey (1959). Gaffey not only devised a consistent estimator  $\hat{g}(\tau)$  for  $g(\tau)$  for the case where the errors of measurement are normally distributed, but also found asymptotic expressions for  $E\hat{g}(\tau)$  and  $\text{Var } \hat{g}(\tau)$ , the bias and sampling variance of his estimator.

Although Gaffey’s estimator is consistent, its convergence to the true  $g(\tau)$  may be slow. If  $\hat{g}(\tau)$  is a frequency distribution, the variance  $\hat{\sigma}_T^2$  [not  $\text{Var } \hat{g}(\tau)$ ] of  $T$  approximately satisfies the equation

$$\hat{\sigma}_T^2 = \sigma_X^2 - c_N \sigma_E^2, \quad (22.5.1)$$

where  $c_N < 1$  is a constant chosen by the statistician.

Since  $\sigma_T^2 = \sigma_X^2 - \sigma_E^2$ , it would seem from (22.5.1) that  $c_N$  should be chosen very close to one; otherwise  $\hat{\sigma}_T^2$  will be seriously biased. However, if  $c_N$  is too close to one, too much weight will be given to the terms of  $\hat{g}(\tau)$  in (22.4.9) that involve estimated fourth- and higher-order derivatives of  $\varphi(x)$ . Since these estimates cannot have any accuracy for samples of the size usually available, it will not be satisfactory to choose  $c_N$  close to one unless  $N$ , the number of

---

\* From Trumpler, R. J., and H. F. Weaver. *Statistical astronomy*. Berkeley: University of California Press, 1953. Used by permission.

observations, is very large. There is a real dilemma here. Gaffey suggests that  $c_N$  be chosen by the formula

$$c_N = 1 - (\log_e N)^{-0.1}. \quad (22.5.2)$$

For a set of data studied by Lord,  $N$  was 388,071, so that

$$c_N = 1 - 1/\sqrt[4^0]{12.87} = 0.225.$$

In (22.5.1), this value of  $c_N$  yields a highly unreasonable estimate of  $\sigma_T^2$ . According to (22.5.2), one would need a sample of roughly  $N = 10^{(1020)}$  observations for  $c_N$  to equal 0.99. The number of elementary particles in the universe is considerably less than  $10^{(102)}$ .

Perhaps some other method of choosing  $c_N$  can produce good results with Gaffey's method. Alternatively, perhaps some modification of Gaffey's estimator can be found for approximating the true  $g(T)$  for samples of, say, 10,000 or 100,000.

## 22.6 The Regression of True Score on Observed Score

For the special case where the conditional distribution of the error of measurement,  $E$ , for given true score is normal with zero mean and fixed variance, there is an elegant mathematical formula for the regression of true score on observed score. If a good, consistent statistical estimator of this regression could be found, it could be used to estimate an examinee's true score from his observed score in those (probably rare) cases where the errors of measurement are conditionally normally distributed.

The derivation given here is essentially the same as that of a formula attributed to Eddington (Trumpler and Weaver, 1953, Section 1.51). The conditional distribution of  $E$  for given true score, and hence the unconditional distribution also, is

$$p(e | \tau) = p(e) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{e^2}{\sigma^2}\right). \quad (22.6.1)$$

Note for future reference that its derivative is

$$\frac{d}{de} p(e) = -\frac{e}{\sigma^2} p(e). \quad (22.6.2)$$

If  $g(\tau)$  is the distribution of  $T$ , the joint distribution of  $T$  and  $E$  is

$$h(\tau, e) = g(\tau)p(e). \quad (22.6.3)$$

Since  $T = X - E$ , the joint distribution of  $X$  and  $E$  is thus

$$f(x, e) = g(x - e)p(e). \quad (22.6.4)$$

If  $\varphi(x)$  is the distribution of  $X$ , then

$$\varphi(x) = \int_{-\infty}^{\infty} g(x - e)p(e) de. \quad (22.6.5)$$

The conditional distribution of  $E$  for given  $x$  is now

$$f(e | x) = \frac{g(x - e)p(e)}{\varphi(x)}. \quad (22.6.6)$$

The conditional mean of  $E$  for given  $x$  is thus

$$\mu(E | x) = \frac{1}{\varphi(x)} \int_{-\infty}^{\infty} eg(x - e)p(e) de. \quad (22.6.7)$$

Integrating by parts and using (22.6.2), we obtain

$$\begin{aligned} \mu(E | x) &= \frac{1}{\varphi(x)} \left[ -\sigma^2 g(x - e)p(e) + \sigma^2 \int p(e)g'_e(x - e) de \right]_{-\infty}^{+\infty} \\ &= \frac{\sigma^2}{\varphi(x)} \int_{-\infty}^{\infty} p(e)g'_e(x - e) de, \end{aligned} \quad (22.6.8)$$

where

$$g'_e(x - e) \equiv \frac{\partial}{\partial e} g(x - e). \quad (22.6.9)$$

The first term in brackets in (22.6.8) vanishes because of the behavior of  $p(e)$  at  $e = \pm\infty$ . The integration by parts is permissible if the last integral in (22.6.8) exists; in particular, if  $g'$  exists and is bounded.

Now

$$\frac{\partial}{\partial x} g(x - e) = -g'_e(x - e). \quad (22.6.10)$$

Substitute this into (22.6.8), and use (22.6.5) to find

$$\mu(E | x) = -\frac{\sigma^2}{\varphi(x)} \frac{\partial}{\partial x} \int_{-\infty}^{\infty} p(e)g(x - e) de = -\sigma^2 \frac{\varphi'(x)}{\varphi(x)}, \quad (22.6.11)$$

where

$$\varphi'(x) \equiv \frac{d}{dx} \varphi(x). \quad (22.6.12)$$

This last step, again, is permissible so long as  $g'$  exists and is bounded.

The conditional mean of the true score for a given observed score, finally, is

$$\mu(T | x) = \mu(X - E | x) = \mu(X | x) - \mu(E | x) = x + \sigma^2[\varphi'(x)/\varphi(x)]. \quad (22.6.13)$$

Thus the regression of true score on observed score is a simple function of the frequency distribution of the observed score, the slope of this distribution, and

the variance of the errors of measurement. If known, this regression can be used to estimate any examinee's true score from his observed score.

Note that if the distribution of  $X$  is unimodal, the estimated true score  $\mu(T | x)$  for any examinee whose observed score is below the mode of the distribution of  $X$  will be higher than his observed score; the estimated true score for any observed score above the mode will be less than the observed score. This effect might be called regression toward the mode. It may be summarized roughly by stating that extreme observed scores should be somewhat discounted as probably attributable in part to extreme errors of measurement.

If the distribution of observed scores is flat within any interval of  $X$ , there will be no regression toward the mode for observed scores in this interval. Note, however, that if the errors are normally distributed with fixed variance  $\sigma^2 > 0$ , and if (22.6.13) is valid, then the distribution of  $X$  cannot be rectangular. We may prove this by supposing the converse. If the distribution were rectangular, then the regression of true score on observed score would be  $\mu(T | x) = x$ , by (22.6.13), and would coincide with the regression of observed score on true score (3.7.1a). However, two linear regressions can coincide only if the two variables ( $X$  and  $T$ ) are perfectly linearly related. This cannot occur if  $\sigma^2 > 0$ .

A formula for  $\mathcal{E}\{[T - \mu(T | x)]^2 | x\}$ , the variance of the errors of prediction when predicting  $T$  from  $x$ , could presumably be derived by an extension of the derivation already given. This is left as an exercise for the interested reader.

In practical work (see Trumpler and Weaver), it is necessary to substitute the first differences of the sample grouped frequency distribution for the derivative required in (22.6.13). To the authors' best knowledge, no one has successfully studied the statistical inference problems involved and found a practical modification of (22.6.13) that will provide a consistent and, if possible, reasonably efficient and unbiased estimator of  $\mu(T | x)$ .

The lack of such an estimator is not too serious if, as is probably often true, the regression  $\mu(T | x)$  is not too nonlinear. In such cases, the linear least-squares regression of Section 3.7 can be used.

## 22.7 The Assumption of Normally Distributed Errors

Pollard (1953) gives a mathematical condition on  $\varphi(x)$ , the frequency distribution of  $X$ , that is both sufficient and necessary for the errors to be normally distributed with constant variance for fixed  $T$ . Since Pollard's condition involves an infinite number of derivatives of  $\varphi(x)$ , it does not readily lend itself to statistical inference. A similar comment applies to a multivariate extension of Pollard's result given by Standish (1956). No adequate alternative practical procedures appear to be available.

Strictly speaking, if the test score  $X$  is the number of items answered correctly, it is clear that  $E$  cannot be normally distributed for fixed  $T$ , since in this case  $X$  is both discrete and bounded. If the test score is neither discrete nor bounded, the assumption of normally distributed errors may be plausible.

At the end of Chapter 10, we summarized a study by Lord (1960) which showed that if the test score is taken as the number of items answered correctly, then the errors of measurement studied were not normally distributed, nor were they distributed independently of true score. In Chapter 23, we shall discuss a possible model for the distribution of the errors of measurement in such cases.

## 22.8 Conditions for Linear Regression of True Score on Observed Score

If (22.6.13) gives the regression of true score on observed score, then the condition for linearity (assuming that  $\sigma^2 \neq 0$ ) is that the last term be a linear function of  $x$ ; that is, the condition is

$$\frac{\varphi'(x)}{\varphi(x)} = A + Bx, \quad (22.8.1)$$

where  $A$  and  $B$  are unknown constants. Integrate both sides to obtain

$$\log \varphi(x) = Ax + \frac{1}{2}Bx^2 + C,$$

or

$$\varphi(x) = \exp(Ax + \frac{1}{2}Bx^2 + C). \quad (22.8.2)$$

Since  $\varphi(x)$  is a frequency distribution, (22.8.2) shows that the observed score  $x$  must be normally distributed. Since the errors are normally distributed, independently of true score, it follows that the true score must be normally distributed also. Thus we have

**Theorem 22.8.1.** *If the errors of measurement are normally distributed independently of true score (as assumed in 22.6), then, under the regularity conditions assumed in Section 22.6, the regression of true score on observed score will be linear if and only if the true scores are normally distributed.*

If a particular population of examinees has a normal distribution of true scores, but a subpopulation is selected in which this distribution is not normal, then the regression of true score on observed score in the subpopulation will be nonlinear.

A general result can be obtained without assuming any particular form for the frequency distribution of the errors of measurement.\* Kendall and Stuart (1961, Section 28.5) have shown that if some variable  $Y_2$  has a linear regression on  $Y_1$ , so that  $\mathcal{E}(Y_2 | Y_1) = A + BY_1$ , then

$$\left. \frac{\partial \psi(\theta_1, \theta_2)}{\partial \theta_2} \right|_{\theta_2=0} = iA + B \frac{d\psi(\theta_1, 0)}{d\theta_1}, \quad (22.8.3)$$

where  $i \equiv \sqrt{-1}$  and  $\psi(\theta_1, \theta_2)$  is the bivariate cumulant generating function of  $Y_1$  and  $Y_2$ .

---

\* The remainder of this section assumes some familiarity with characteristic functions. The reader may skip to the next section without loss of continuity.

This result can be applied to the present problem by starting with the bivariate characteristic function of  $X$  and  $T$ , which may be written

$$\begin{aligned}\Phi(\theta_1, \theta_2) &\equiv \mathcal{E} \exp(i\theta_1 X + i\theta_2 T) = \mathcal{E} \exp[i\theta_1(T + E) + i\theta_2 T] \\ &= \mathcal{E}\{\exp[i(\theta_1 + \theta_2)T] \exp(i\theta_1 E)\} = \Phi_T(\theta_1 + \theta_2)\Phi_E(\theta_1),\end{aligned}\quad (22.8.4)$$

where  $\Phi_T(\theta)$  and  $\Phi_E(\theta)$  are the characteristic functions of true score and error, respectively. Thus, the cumulant generating function for observed score and true score can be written

$$\psi(\theta_1, \theta_2) = \psi_T(\theta_1 + \theta_2) + \psi_E(\theta_1). \quad (22.8.5)$$

Differentiate (22.8.5) to obtain

$$\begin{aligned}\frac{\partial \psi(\theta_1, \theta_2)}{\partial \theta_1} &= \frac{d\psi_T(\theta)}{d\theta} \Big|_{\theta=\theta_1+\theta_2} + \frac{d\psi_E(\theta)}{d\theta} \Big|_{\theta=\theta_1}, \\ \frac{\partial \psi(\theta_1, \theta_2)}{\partial \theta_2} &= \frac{d\psi_T(\theta)}{d\theta} \Big|_{\theta=\theta_1+\theta_2}.\end{aligned}$$

Insert these results into (22.8.3) to obtain a necessary condition for linearity of regression:

$$\frac{d\psi_T(\theta)}{d\theta} \Big|_{\theta=\theta_1} = iA + B \left[ \frac{d\psi_T(\theta)}{d\theta} + \frac{d\psi_E(\theta)}{d\theta} \right]_{\theta=\theta_1}.$$

Without loss of generality, we can suppose that a constant has been subtracted from all observed scores so that the mean of  $X$  in the population of examinees is zero. Thus  $A = 0$  when  $Y_2$  = true score and  $Y_1$  = observed score. Also, by (3.7.2),  $B = \rho$ , where  $\rho$  is the test reliability coefficient, which we assume to be nonzero. The last displayed equation can now be written

$$(1 - \rho) \frac{d\psi_T(\theta)}{d\theta} = \rho \frac{d\psi_E(\theta)}{d\theta}. \quad (22.8.6)$$

Integrate both sides to obtain, finally,

$$(1 - \rho)\psi_T(\theta) = \rho\psi_E(\theta). \quad (22.8.7)$$

(Any constants of integration must cancel out, since  $\psi(0) = 0$  for any random variable.) Thus, *when the errors are distributed independently of true score, a necessary condition for true score to have a linear regression on observed score is that  $\psi_T(\theta)$ , the cumulant generating function of the true scores, be a constant multiple of  $\psi_E(\theta)$ , the cumulant generating function of the errors of measurement.*

If the cumulant generating function can be expanded in a power series, then

$$\frac{d^r \psi(\theta)}{d\theta^r} \Big|_{\theta=0}$$

is the  $r$ th cumulant. Thus, under regularity conditions, (22.8.7) shows that

$$(1 - \rho)\kappa_T^{(r)} = \rho\kappa_E^{(r)}, \quad r = 2, 3, \dots. \quad (22.8.8)$$

Taken together with (22.2.4), this shows that the true-score cumulants after the first are the same as the observed-score cumulants, except for the constant factor  $\rho$ :

$$\kappa_T^{(r)} = \rho\kappa_X^{(r)}, \quad r = 2, 3, \dots. \quad (22.8.9)$$

A special case satisfying (22.8.7), (22.8.8), and (22.8.9) is, of course, the case where the true scores and the errors of measurement are both normally distributed. Here

$$\psi_T(\theta) = -\frac{1}{2}\sigma_T^2\theta^2, \quad \psi_E(\theta) = -\frac{1}{2}\sigma_E^2\theta^2. \quad (22.8.10)$$

The reader may verify (22.8.8) and (22.8.7) for this special case, remembering that  $\rho = \sigma_T^2/\sigma_X^2$  and  $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$ .

## 22.9 Conditions for Linear Multiple Regression of True Score on Two or More $\tau$ -Equivalent Observed Scores

Ferguson (1955, Theorem 5) has proved a theorem that can be specialized for present purposes.

**Theorem 22.9.1.** *Let the random variables  $X_1, X_2, \dots, X_g, \dots, X_n$ ,  $n \geq 2$ , be  $\tau$ -equivalent measurements on a population of examinees, the errors of measurement  $X_g - T$ ,  $g = 1, 2, \dots, n$ , being distributed independently of each other and of the true score  $T$  with zero means and nonzero variances. Then for the multiple regression of  $T$  on  $X_1, \dots, X_n$  to be linear, it is necessary and sufficient that  $T, X_1, \dots, X_n$  be jointly normally distributed.*

The proof of the theorem will not be given here.

It appears from this theorem that strict linear regression of true score on observed score is a rather specialized and unusual situation.

## 22.10 Conditions for Linear Regression of One Measurement on Another

Lindley (1947, Section 3.1) has proved a theorem that includes the results of Section 22.8 as special cases. Rephrased for present purposes, it is

**Theorem 22.10.1.** *Given that*

- a) *true score  $T_1$  has a linear regression on true score  $T_2$  with slope  $\beta$ , and*
  - b) *the errors of measurement  $E_1 = X_1 - T_1$  and  $E_2 = X_2 - T_2$  are distributed independently of each other and of  $T_1$  and  $T_2$ ,*
- then a necessary and sufficient condition for the regression of observed score  $X_1$  on observed score  $X_2$  to be linear is that the cumulant generating function of  $T_2$  be a multiple of the cumulant generating function of  $E_2$ . Specifically, this*

condition is

$$(\beta - B)\psi_{T_2}(\theta) = B\psi_{E_2}(\theta), \quad (22.10.1)$$

where  $B$  is the slope of the regression of  $X_1$  on  $X_2$ .

A proof of this result is also given by Kendall and Stuart (1961, Section 29.57).

Lindley has given several related theorems, including the multivariate generalization of Theorem 22.10.1. The interested reader is referred to his article and to Ferguson's.

It appears from this theorem that strict linear regression of one observed score on another is a specialized occurrence. However, approximate linear regression does seem to hold in practice for many sets of empirical data.

### Exercise

- 22.1. Suppose  $X$  and  $Y$  are essentially  $\tau$ -equivalent. Show that the regression of  $X$  on  $Y$  is linear if and only if the regression of true score on observed score is linear for  $Y$ .

### References and Selected Readings

- EDDINGTON, A. S., On a formula for correcting statistics for the effects of a known probable error of observation. *Royal Astronomical Society Monthly Notices*, 1913, **73**, 359–360.
- FERGUSON, T., On the existence of linear regression in linear structural relations. *University of California Publications in Statistics*, 1955, **2**, No. 7.
- GAFFEY, W. R., A consistent estimator of a component of a convolution. *Annals of Mathematical Statistics*, 1959, **30**, 198–205.
- GIRSHICK, M. A., and L. J. SAVAGE, Bayes and minimax estimates for quadratic loss functions. In J. Neyman (Ed.), *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. Berkeley: University of California Press, 1951, pp. 53–73.
- HIRSCHMAN, I. I., and D. V. WIDDER, *The convolution transform*. Princeton, N.J.: Princeton University Press, 1955.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. Vol. 1: *Distribution theory*. New York: Hafner, 1958.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. Vol. 2: *Inference and relationship*. New York: Hafner, 1961.
- KURTH, R., On Eddington's solution of the convolution integral equation. *Rendiconti del Circolo Matematico di Palermo*, 1965, **14**, 76–84.
- LINDLEY, D. V., Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society*, 1947, **9**, 218–244.
- LORD, F. M., An empirical study of the normality and independence of errors of measurement in test scores. *Psychometrika*, 1960, **25**, 91–104.

- NOVICK, M. R., The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, **3**, 1-18.
- NOVICK, M. R., and C. LEWIS, Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 1967, **32**, 1-13.
- PATIL, G. P. (Ed.), *Classical and contagious distributions*. New York: Pergamon Press, 1965.
- PATIL, G. P., and S. W. JOSHI, Bibliography of classical and contagious discrete distributions. ARL 66-0185, Aerospace Research Laboratories, U.S. Air Force, September 1966.
- PITMAN, E. J. G., The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika*, 1939, **30**, 391-421.
- POLLARD, H., Distribution functions containing a Gaussian factor. *Proceedings of the American Mathematical Society*, 1953, **4**, 578-582.
- ROBBINS, H., An empirical Bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Vol. 5. Berkeley: University of California Press, 1956, pp. 157-163.
- ROBBINS, H., The empirical Bayes approach to testing statistical hypotheses. *Review of the International Statistical Institute*, 1963, **31**, 195-208.
- ROBBINS, H., The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 1964, **35**, 1-20.
- SCARBOROUGH, J. B., *Numerical mathematical analysis*, 3rd ed. Baltimore: The Johns Hopkins Press, 1955.
- STANDISH, C., *N*-dimensional distributions containing a normal component. *Annals of Mathematical Statistics*, 1956, **27**, 1161-1165.
- TEICHER, H., On the mixture of distributions. *Annals of Mathematical Statistics*, 1960, **31**, 55-73.
- TRUMPLER, R. J., and H. F. WEAVER, *Statistical astronomy*. Berkeley: University of California Press, 1953.

## CHAPTER 23

# BINOMIAL ERROR MODELS

### 23.1 Introduction

The present chapter summarizes a strong true-score theory (Lord, 1965) that is based on the assumption that the conditional distribution of observed score for given true score is binomial or compound binomial. In many ways this chapter parallels the preceding one, but here the error distribution is not assumed to be independent of true score.

In Sections 2 and 4, we shall outline the basic assumptions of the binomial model. In Section 5, we shall derive equations for the (possibly nonlinear) regression of true score on observed score. This regression can be estimated rather accurately for large groups of examinees and can be used to estimate each examinee's true score from his observed score.

If this regression is linear, and if the binomial model holds, then the observed-score distribution must be a negative hypergeometric distribution, as shown in Section 6. This distribution can be of practical use in fitting or smoothing actual distributions of observed scores. It may well be preferable to other currently available techniques.

Under the binomial model, regardless of linearity of regression, the first  $n$  moments of the true-score distribution can be determined from the moments of the observed-score distribution. The same is true of the first  $n$  bivariate moments of the joint distribution of observed scores and true scores. This bivariate distribution provides complete information about the relation of observed measurements to true measurements. These results appear in Section 8.

In Sections 8 and 9, we shall discuss some implications of the binomial model that lead to a correction involving substitution of a compound binomial distribution for the simple binomial. Section 10 outlines this new model.

In Section 11, we shall very briefly develop some of the implications of the model. The model can be used

- 1) to estimate the effect of using a fallible measure for selection purposes;
- 2) to obtain approximate nationwide test norms distributions without administering the test to the norms group;
- 3) to describe the properties of the test as a measuring instrument;
- 4) for other purposes.

The full utilization of this model depends on obtaining the solution to an integral equation. Although various methods are available for this purpose, almost all ignore the problems of statistical inference from fallible data. Thus there are several questions that must be investigated before the model is ready for routine application to practical problems. For this reason, the practical application of the model is not extensively illustrated in the exposition here.

### 23.2 Definitions and Assumptions

In this chapter, we shall be concerned with tests on which  $x_a$ , the total score of examinee  $a$ , is the number of items answered correctly:

$$x_a = \sum_{i=1}^n u_{ia}, \quad (23.2.1)$$

where  $u_{ia} = 0$  or  $1$  is the examinee's score on item  $i$ . We denote by  $X$  the random variable over examinees corresponding to  $x_a$ . In this chapter, there is no explicit consideration of replicate measurements.

We adopt the assumption, developed in Chapter 2, that for fixed true score  $\xi$ , the error of measurement defined by

$$E \equiv X - \xi \quad (23.2.2)$$

is unbiased; that is,

$$\mathcal{E}(E | \xi) = 0, \quad (23.2.3)$$

where the expectation is over examinees. It follows, of course, that

$$\mathcal{E}(X | \xi) = \xi. \quad (23.2.4)$$

Since  $x$  is the number of right answers, it is clear that  $0 \leq x \leq n$ . From (23.2.4), it follows that  $0 \leq \xi \leq n$ .

If  $\text{Prob}(X > 0 | \xi = 0)$  were nonzero, then  $\mathcal{E}(X | \xi = 0)$  would be nonzero, which would contradict (23.2.4). It follows that examinees with  $\xi = 0$  always (with probability one) get a score of  $x = 0$ . Similarly, examinees with  $\xi = 1$  always get a score of  $x = 1$ . This conclusion shows that *under any model with bounded observed score and unbiased errors (not all zero), the conditional distribution of the observed score cannot be independent of true score; equally, the conditional distribution of the error of measurement cannot be independent of true score*. This contradicts Assumption 22.1.1, which was basic to everything in the preceding chapter. We pointed out there that this assumption is not completely plausible when observed score is the number of right answers. Thus we must now go on to consider cases other than those treated in Chapter 22.

Note that  $X$  can only assume integral values, whereas  $\xi$  is presumably a continuous variable. Thus the conditional distribution of  $X$  for given  $\xi$  must be a discrete distribution depending on the continuous parameter  $\xi$ . A plausible distributional form is the binomial. In Section 11.9, we found this same distribution as an outcome of the item-sampling model for binary items.

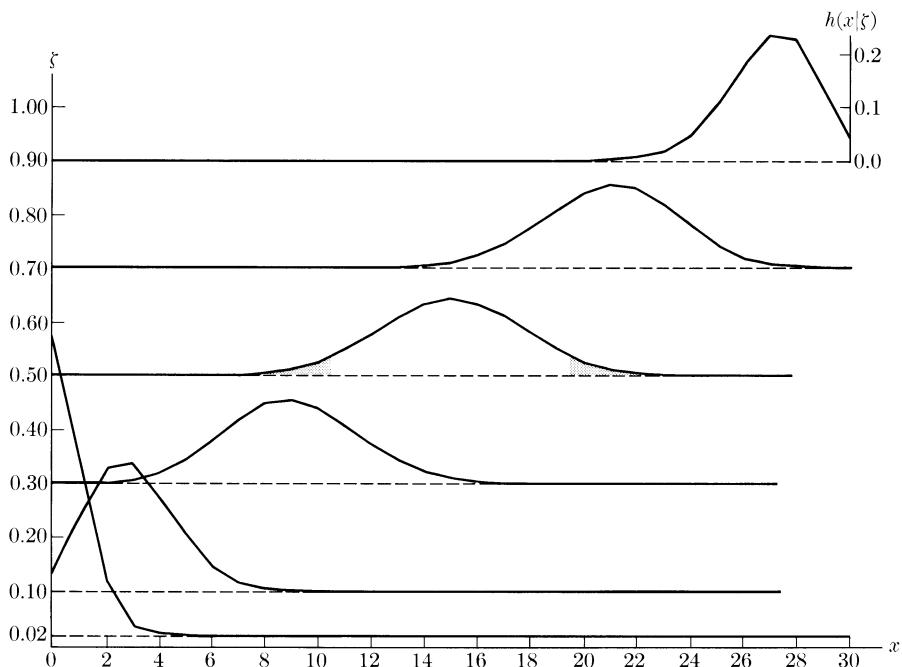


FIG. 23.2.1. Binomial conditional distributions of observed score  $X$  for various true scores  $\xi$  when  $n = 30$ .

The first sections of the present chapter, then, will be based on

**Assumption 23.2.1.** *The conditional distribution of observed score  $X$  for given (proportion-correct) true score  $\xi \equiv \xi/n$  is the binomial distribution*

$$h(x | \xi) = \binom{n}{x} \xi^x (1 - \xi)^{n-x}, \quad x = 0, 1, \dots, n. \quad (23.2.5)$$

Figure 23.2.1 shows the conditional distribution (frequency polygon) of  $X$  for several values of  $\xi$ .

The symbol  $\xi$  is used here for the proportion-correct true score because of the close relation to the formulas already given in Section 11.9. It should not be assumed that this  $\xi$  has all the characteristics of the true score in the item-sampling model, nor of the generic true score of Chapters 8 and 9. Actually, (23.2.5) is the only assumption necessary for the developments here relating to a single test form; the term “true score” need not be defined here other than by (23.2.5).

On the other hand, (23.2.5) is consistent with the basic assumptions of the classical model for a single test given by (3.1.2) and (3.1.3):

$$\mathcal{E}E = 0, \quad \rho_{\xi E} = 0.$$

Thus (23.2.5) is also consistent with all the results that have been deduced solely from these assumptions. These matters will be considered in detail in Section 23.9, after we have derived the necessary formulas. Binomial models for two or more tests are briefly discussed in Section 23.11.

### 23.3 Confidence Intervals for True Score

Tables of the binomial frequency distribution show that under the binomial error model, approximately 90.1% of all examinees with a true score of 0.50 will score more than 10 and less than 20 right answers on a 30-item test. This 90.1% interval is indicated by the unshaded portion of the polygon for  $\xi = 0.50$  in Fig. 23.2.1. Since  $x$  assumes only integral values, it is impossible to find an interval containing exactly 90% of the observed scores.

Similar intervals containing close to 90%, but always at least 90%, of the observed scores can be set up for any  $\xi$  in the range  $0 \leq \xi \leq 1$ . In Figure 23.3.1, such intervals are marked off for  $\xi = 0.51, 0.52, 0.53, \dots, 1.00$ . In the lower half of the same figure, the points marking off the intervals have been connected to indicate (approximate) 90% intervals for every possible value of  $\xi$  from 0 to 0.50.

Because  $x$  is integer-valued, the boundaries of the intervals are irregular. In general, it is impossible to construct an interval with equal frequencies in

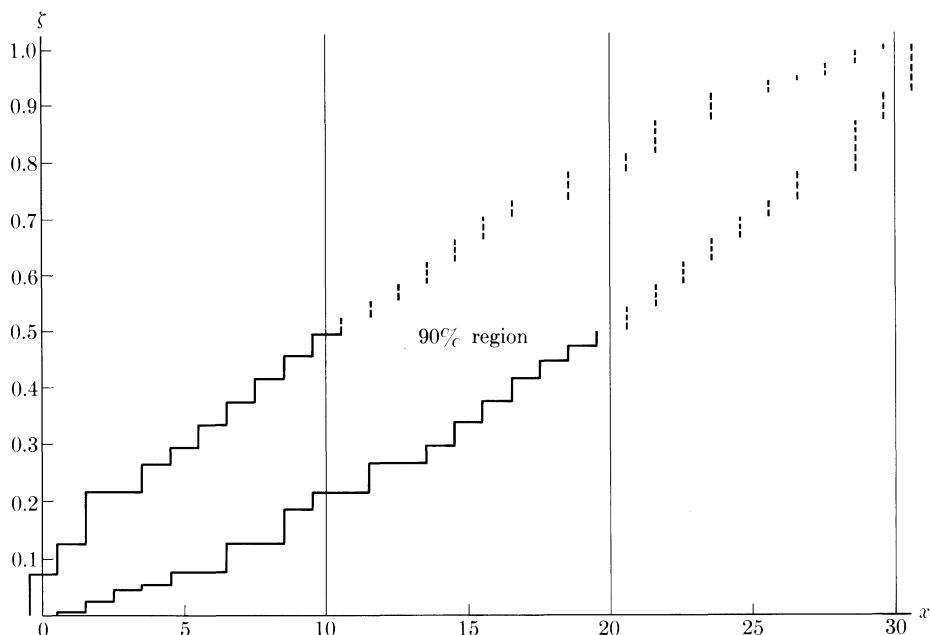


FIG. 23.3.1. Ninety % confidence region for estimating an examinee's true score from his observed score.

each tail. Even the choice of the interval is somewhat arbitrary, within the requirement that it contain 90% or more of the observed scores for fixed  $\xi$ . Intervals that are in some sense optimal are given by Blyth and Hutchinson (1960), by Crow (1956), and by other authors referenced therein.

Now clearly, for a randomly chosen examinee, no matter what his true score may be, the probability is (at least) 90% that his true score will fall in the area labeled "90% region". This follows from the way the boundaries of the region have been constructed. Suppose, again, that a randomly chosen examinee obtains a score of  $x = 10$  on the 30-item test. If we assert that his true score falls in the 90% region, we shall have (at least) a 90% chance of being correct, in the sense that in the long run, similarly based assertions about randomly chosen examinees will be correct (at least) 90% of the time. But since his observed score is 10, the assertion that his true score falls in the 90% region is the same (within the limits of accuracy of the chart) as the assertion that his true score lies in the interval  $0.21 < \xi < 0.50$ ; we see this by reading vertically upward from  $x = 10$  in the figure. Thus, for this examinee, the interval  $0.21 < \xi < 0.50$  is an (at least) 90% confidence interval for  $\xi$ .

If a second randomly chosen examinee obtains an observed score of  $x = 2$ , we see from the figure that the interval  $0.01 < \xi < 0.13$  is a 90% confidence interval for him. *If we assert that each randomly chosen examinee has a true score within the confidence interval similarly derived from his observed score, then we shall be correct 90% of the time.*

It is important to note that *no confidence statement can be made about a particular, nonrandomly chosen examinee in whom we happen to be interested. Nor can any confidence statement be made about those examinees who have some specified observed score  $x = x_0$ .* It might seem natural to assert that 90% of the time, examinees with observed scores  $x = 10$  will have true scores that fall between 0.21 and 0.50. This statement cannot be made with any confidence, however: In any infinite population of examinees, there may conceivably be not a single examinee with a true score lying between 0.21 and 0.50, but some examinees from the infinite population may still obtain scores of  $x = 10$ . In such a case, the assertion that the examinee's true score lies in the usual confidence interval is incorrect for every examinee scoring  $x = 10$ . This assertion is still correct 90% of the time, however, for examinees chosen at random from the population of examinees.

### 23.4 The Fundamental Equation

For the binomial error model, the basic equation (22.3.3) now becomes

$$\varphi(x) = \binom{n}{x} \int_0^1 g(\xi) \xi^x (1 - \xi)^{n-x} d\xi, \quad x = 0, 1, \dots, n, \quad (23.4.1)$$

where  $\varphi(x)$  is the distribution of observed scores and  $g(\xi)$  is the unknown distribution of true scores. Note that  $\xi$  is a proportion and thus ranges from 0 to 1.

If  $g(\xi)$  can be estimated, then so can the bivariate distribution of observed score and true score. This latter distribution describes the relation of observed score to true score completely, making it possible to estimate the true score of any examinee from his observed score.

We shall discuss a number of implications of (23.4.1). The results obtained will frequently parallel those of the preceding chapter.

### 23.5 Regression of True Score on Observed Score

Equation (23.2.4) holds under the binomial assumption (23.2.5). Therefore it is clear that the regression of observed score on true score is linear. For the proportion-correct true score, this regression is given by

$$\mathcal{E}(X | \xi) = n\xi. \quad (23.5.1)$$

The other regression is the one needed to infer an examinee's true score from his observed score. As in Section 3.7, the linear regression of true score on observed score can be written explicitly if the linear regression coefficient  $\beta_{\xi|x} = \sigma_\xi \rho_{\xi|x}/\sigma_x$  is known. This last can be determined from (23.8.6) and (23.8.3).

If the regression is not really linear, however, it will be better to have the formula for the nonlinear regression of true score on observed score. The derivation of the desired formula is the subject of the present section.

By the basic equation (23.2.5), the conditional distribution of  $\xi$  for given  $x$  is

$$p(\xi | x) = \frac{g(\xi)h(x | \xi)}{\varphi(x)} = \frac{1}{\varphi(x)} g(\xi) \binom{n}{x} \xi^x (1 - \xi)^{n-x}, \quad (23.5.2)$$

and the regression of  $\xi$  on  $x$  is

$$\mu_{\xi|x} = \frac{1}{\varphi(x)} \binom{n}{x} \int_0^1 g(\xi) \xi^{x+1} (1 - \xi)^{n-x} d\xi.$$

Subtract each side of this equation from unity, clear of fractions, use (23.4.1), and combine terms to obtain

$$\begin{aligned} \varphi(x)(1 - \mu_{\xi|x}) &= \int_0^1 g(\xi) \binom{n}{x} \xi^x (1 - \xi)^{n-x} d\xi - \int_0^1 g(\xi) \binom{n}{x} \xi^{x+1} (1 - \xi)^{n-x} d\xi \\ &= \int_0^1 g(\xi) \frac{n!}{x!(n-x)!} \xi^x (1 - \xi)^{n-x+1} d\xi \\ &= \frac{n-x+1}{x} \int_0^1 g(\xi) \frac{n!}{(x-1)!(n-x+1)!} \xi^x (1 - \xi)^{n-x+1} d\xi \\ &= \frac{n-x+1}{x} \varphi(x-1) \mu_{\xi|x-1}, \quad x = 1, 2, \dots, n. \end{aligned} \quad (23.5.3)$$

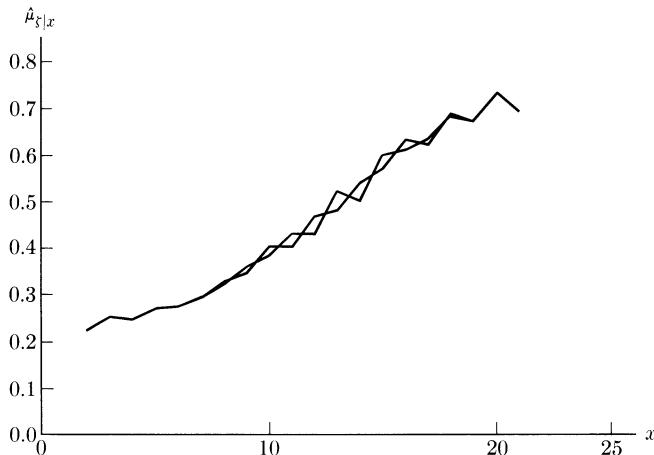


FIG. 23.5.1. Two estimates of the regression of true score  $\xi$  on observed score  $x$  ( $N = 4000$ ). [From F. M. Lord, An approach to mental test theory. *Psychometrika*, 1959, **24**, 283–302. Used by permission.]

Thus, finally, the mean true score for any permissible value of  $x$  except  $x = 0$  can be obtained from  $\varphi(x)$ ,  $\varphi(x - 1)$ , and the mean true score at  $x - 1$  by the equation

$$\mu_{\xi|x} = 1 - \frac{n - x + 1}{x} \frac{\varphi(x - 1)}{\varphi(x)} \mu_{\xi|x-1}, \quad x = 1, 2, \dots, n. \quad (23.5.4)$$

Equation (23.5.4) really represents  $n$  independent linear equations in the  $n + 1$  unknowns  $\mu_{\xi|0}$ ,  $\mu_{\xi|1}$ ,  $\mu_{\xi|2}$ , ...,  $\mu_{\xi|n}$ . If one more mathematically independent equation involving no new unknowns could be found, the regression would be completely determined. However, it can be shown by counterexamples (not presented here) that the regression of true score on observed score is not uniquely determined by the observed-score distribution.

If for some value, say  $x_0$ , of  $X$ , one chooses an arbitrary value of  $\mu_{\xi|x_0}$ , Eq. (23.5.4) then yields a unique “regression” of true score on observed score. Most such regressions are jagged, but for some choice of  $\mu_{\xi|x_0}$  the regression will be most nearly smooth, in some specified sense. Experience has shown, as we shall illustrate below, that the requirement that the regression be smooth effectively eliminates the indeterminacy. This seems to be a reasonable procedure for sufficiently large samples. The extent and nature of the sampling error of an estimated regression determined in this fashion have not been investigated, and consequently the necessary sample size cannot be specified here. Except for very large samples, some preliminary smoothing of the observed-score distribution is usually necessary to avoid unpleasant irregularities in the estimated regression.

Figure 23.5.1 shows two estimates of the regression of true score on observed score for a 25-item vocabulary test. These estimates were obtained by applying (23.5.4) to the (smoothed) observed-score distribution of  $N = 4000$  examinees. The two estimates were purposely made as different as they might be and still be monotonic increasing for  $x = 10, 11, 12$ . For an examinee with an observed score  $x = 2$ , the estimated true score is 0.22. For an examinee with observed score  $x = 12$ , the estimated true score lies between 0.43 and 0.47.

If smooth regression estimates had been required, the two estimates would have been even more alike. Any such estimated regression obtained from these data will show the same slight curvature that is shown in the figure. The indeterminacy in Eq. (23.5.4) cannot be used to make the estimated regression appear linear.

The problem of estimating the regression of true score on observed score is an *empirical Bayes estimation problem*. An approximate method for doing this, even in moderate-sized samples, has been outlined by Maritz (1966).

### 23.6 Negative Hypergeometric Distribution of Test Scores

Under what conditions is the regression function (23.5.4) linear? If it is linear, then it is identical with the linear minimum mean squared error regression function, and may be written

$$\mu_{\xi|x} = \mu_\xi + \beta_{\xi X}(x - \mu_X), \quad (23.6.1)$$

where  $\beta_{\xi X} = \sigma_\xi \rho_{\xi X} / \sigma_X$ . From (23.5.1), it is clear that

$$\mu_X = n\mu_\xi; \quad (23.6.2)$$

hence the linear regression formula can be written

$$\mu_{\xi|x} = \beta_{\xi X}x + \mu_\xi(1 - n\beta_{\xi X}). \quad (23.6.3)$$

Similarly

$$\mu_{\xi|x-1} = \beta_{\xi X}(x - 1) + \mu_\xi(1 - n\beta_{\xi X}).$$

Substitute these last two equations into (23.5.3) to obtain

$$\begin{aligned} & \varphi(x)[1 - \beta_{\xi X}x - \mu_\xi(1 - n\beta_{\xi X})] \\ &= \frac{n-x+1}{x} \varphi(x-1)[\beta_{\xi X}(x-1) + \mu_\xi(1 - n\beta_{\xi X})]. \end{aligned}$$

Replacing  $\varphi(x)$  by  $h(x)$  to distinguish this special distribution of  $x$  from the general case, we may write this last result as

$$h(x) = \frac{n-x+1}{x} \frac{a-1+x}{b+1-x} h(x-1) \quad (23.6.4)$$

with the aid of the new symbols

$$a = (-n + 1/\beta_{\xi}x)\mu_{\xi}, \quad (23.6.5)$$

$$b = -a - 1 + 1/\beta_{\xi}x. \quad (23.6.6)$$

In (23.6.15) and (23.6.16), we shall see that  $a > 0$  and  $b > n - 1$ .

From (23.6.4),

$$h(1) = \frac{na}{b} h(0) \quad \text{and} \quad h(2) = \frac{n-1}{2} \frac{a+1}{b-1} h(1) = \frac{n(n-1)a(a+1)}{2b(b-1)} h(0).$$

In general,

$$h(x) = \frac{n^{[x]}(a)_x}{b^{[x]}x!} h(0) = \frac{(-n)_x(a)_x}{(-b)_x x!} h(0), \quad x = 0, 1, \dots, n, \quad (23.6.7)$$

where

$$\begin{aligned} n^{[x]} &\equiv n(n-1) \cdots (n-x+1), & n^{[0]} &\equiv (a)_0 \equiv 1, \\ (a)_x &\equiv a(a+1) \cdots (a+x-1). \end{aligned}$$

The probability generating function of  $h(x)$  is proportional to the hypergeometric function

$$F(-n, a; -b; t) \equiv \sum_{x=0}^{\infty} \frac{(-n)_x(a)_x}{(-b)_x x!} t^x. \quad (23.6.8)$$

This is a terminating series when  $n$  is a positive integer. The distribution  $h(x)$  with  $a, b > 0$  is called *negative hypergeometric*. The hypergeometric function with  $a, b < 0$  is the probability generating function for the distribution of the number of white balls obtained in  $N$  random drawings *without replacement* from a *finite* population of black and white balls. This last distribution is usually referred to simply as the *hypergeometric distribution*. It has a variance that is always less than that of the binomial distribution with the same mean; the negative hypergeometric, as we shall see, has a variance that is always greater than this.

Since  $F(-n, a; -b; 1) = (a+b)^{[n]}/b^{[n]}$  [Erdelyi, 1953, Vol. 1, Eqs. 2.8 (46) and 1.2 (4)], and since  $\sum_0^n h(x) = 1$ , we find by summing (23.6.7) on  $x$  that

$$h(0) = b^{[n]} / (a+b)^{[n]}. \quad (23.6.9)$$

Thus, finally, we have the result that *under the binomial error model, if true score has a linear regression on observed score, then the observed score has the negative hypergeometric distribution*

$$h(x) \equiv \frac{b^{[n]}}{(a+b)^{[n]}} \frac{(-n)_x(a)_x}{(-b)_x x!}, \quad x = 0, 1, \dots, n, \quad (23.6.10)$$

where  $n$  is the number of test items and  $a$  and  $b$  are parameters to be determined.

The converse is also true: *Under the binomial error model, if the observed-score distribution is negative hypergeometric, then the regression of true score on observed score is linear.*

Standard formulas (Kendall and Stuart, 1958, Eq. 5.55) for the moments of any hypergeometric distribution show that for the negative hypergeometric,

$$\mu_X = na/D, \quad (23.6.11a)$$

$$\sigma_X^2 = \frac{na}{D} \frac{D-a}{D} \frac{D+n}{D+1}, \quad (23.6.11b)$$

where

$$D \equiv a + b - n + 1. \quad (23.6.12)$$

Equations (23.6.11) can be solved to express  $a$  and  $b$  in terms of the moments of the observed-score distribution:

$$a = (-1 + 1/\alpha_{21})\mu_X, \quad (23.6.13a)$$

$$b = -a - 1 + n/\alpha_{21}, \quad (23.6.13b)$$

where

$$\alpha_{21} \equiv \frac{n}{n-1} \left[ 1 - \frac{\mu_X(n-\mu_X)}{n\sigma_X^2} \right]. \quad (23.6.14)$$

This last quantity is the same as the Kuder-Richardson formula 21 reliability coefficient, previously derived as (4.4.11).

We are now able to find the ranges of the parameters  $a$  and  $b$ . In (23.8.12), it will be shown that  $\alpha_{21} = \rho^2(X, \zeta)$ . If we leave aside the degenerate cases where the error variance is zero or infinite, we may conclude that  $0 < \alpha_{21} < 1$ , and hence, by (23.6.13a), that  $a > 0$ . Since  $0 < \mu_X < n$ , (23.6.11a) shows that

$$0 < a < D. \quad (23.6.15)$$

Finally, substituting (23.6.12) into (23.6.15), we obtain

$$b > n - 1. \quad (23.6.16)$$

Both  $a$  and  $b$  can be indefinitely large.

The negative hypergeometric distribution is recommended for graduating (fitting) observed distributions of number-right scores. A practical procedure is to compute the mean and variance of the sample distribution and then substitute these for  $\mu_X$  and  $\sigma_X^2$  in (23.6.14) and (23.6.13) to obtain estimates of the parameters  $a$  and  $b$ . The second fraction in (23.6.10) can then be readily computed for  $x = 0, 1, \dots, n$ , successively. The first fraction in (23.6.10) is equal to the reciprocal of the sum over  $x$  of all the second fractions just computed.

The numerical example given below illustrates the process for a very simple case that avoids the use of decimals. We start with an observed-score distribution for a 5-item test with  $\mu_X = 2$  and  $\alpha_{21} = 0.5$ . From (23.6.13), we find

that  $a = 2$  and  $b = 7$ . Table 23.6.1 shows the values that the second fraction in (23.6.10) assumes when  $x = 0, 1, \dots, 5$ .

Table 23.6.1

$x$	$\frac{(-n)_x(a)_x}{(-b)_x x!}$
0	$\frac{1 \times 1}{1 \times 1} = \frac{7}{7}$
1	$\frac{7}{7} \times \frac{-5 \times 2}{-7 \times 1} = \frac{10}{7}$
2	$\frac{10}{7} \times \frac{-4 \times 3}{-6 \times 2} = \frac{10}{7}$
3	$\frac{10}{7} \times \frac{-3 \times 4}{-5 \times 3} = \frac{8}{7}$
4	$\frac{8}{7} \times \frac{-2 \times 5}{-4 \times 4} = \frac{5}{7}$
5	$\frac{5}{7} \times \frac{-1 \times 6}{-3 \times 5} = \frac{2}{7}$
Total	$\frac{42}{7} = 6$

The rightmost column is proportional to  $h(x)$ . Since the sum of this column is six, the distribution  $h(x)$  is

$x$	0	1	2	3	4	5
$h(x)$	$\frac{7}{42}$	$\frac{10}{42}$	$\frac{10}{42}$	$\frac{8}{42}$	$\frac{5}{42}$	$\frac{2}{42}$

Figure 23.6.1 shows some negative hypergeometric distributions that have been fitted to test data by the method of moments, as just described. The fits are less than perfect, but probably better than would be obtained by other methods requiring only the mean and variance of the observed-score distribution.

The agreement between model and data can also be checked against the following theoretical result. Suppose there are two statistically identical test forms  $X$  and  $Y$  having  $n$  items each. Also suppose that the scores  $X$  and  $Y$  are independently distributed when  $\xi$  is fixed (this is the assumption of local independence, defined in Section 16.3), so that the trivariate distribution of  $X$ ,  $Y$ , and  $\xi$  is

$$f(x, y, \xi) = g(\xi) \binom{n}{x} \xi^x (1 - \xi)^{n-x} \binom{n}{y} \xi^y (1 - \xi)^{n-y}. \quad (23.6.17)$$

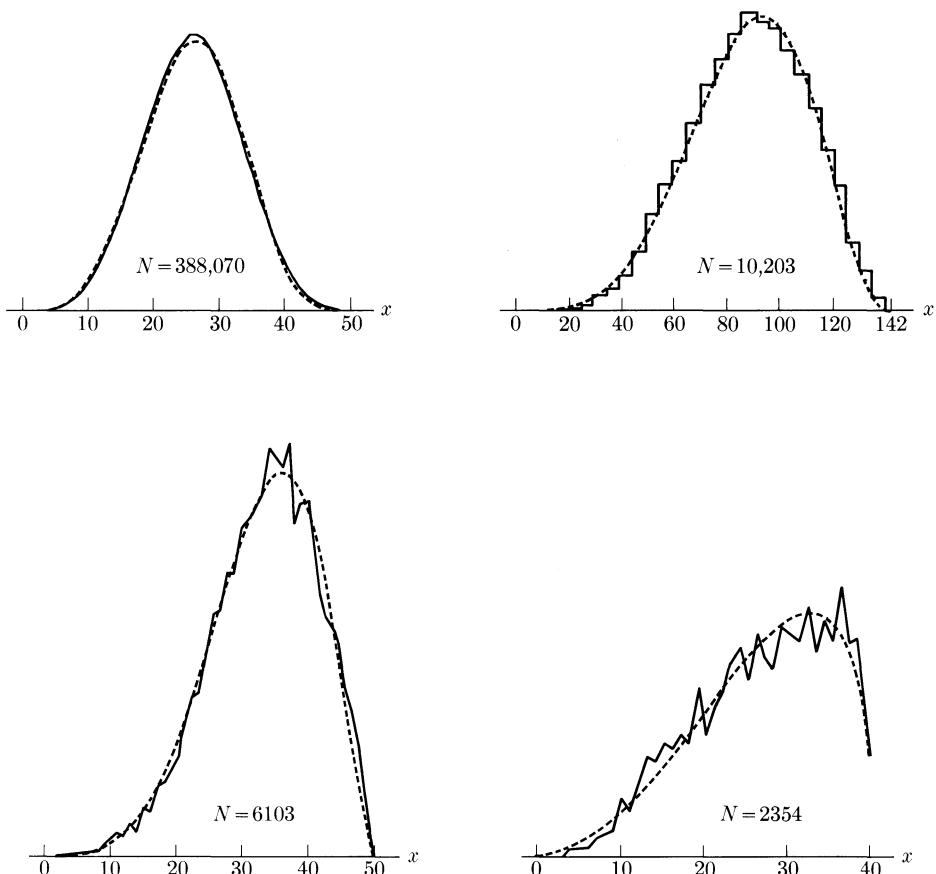


FIG. 23.6.1. Negative hypergeometric distributions (----) fitted to four sets of test-score data. [From J. A. Keats, and F. M. Lord, A theoretical distribution for mental test scores. *Psychometrika*, 1962, **27**, 59-72. Used by permission.]

Denote the score on all  $2n$  items by  $S \equiv X + Y$ . Then the trivariate distribution of  $X$ ,  $S$ , and  $\xi$  is

$$\varphi(x, s, \xi) = g(\xi) \binom{n}{x} \binom{n}{s-x} \xi^s (1 - \xi)^{2n-s}. \quad (23.6.18)$$

The bivariate distribution of  $S$  and  $\xi$  is found from this to be

$$f(s, \xi) = g(\xi) \xi^s (1 - \xi)^{2n-s} \sum_{x=0}^n \binom{n}{x} \binom{n}{s-x}. \quad (23.6.19)$$

It can be shown that the sum is equal to  $\binom{2n}{s}$ .

Now divide (23.6.18) by (23.6.19) to obtain the conditional distribution of  $X$  for given  $S$  and  $\xi$ :

$$f(x | s, \xi) = \binom{n}{x} \binom{n}{s-x} / \binom{2n}{s}. \quad (23.6.20)$$

But this result does not depend on  $\xi$ , nor on the unknown distribution  $g(\xi)$ . Consequently the conditional distribution of  $X$  for fixed  $S$ , regardless of the distribution of true score, is

$$f(x | s) = \binom{n}{x} \binom{n}{s-x} / \binom{2n}{s}. \quad (23.6.21)$$

This is a positive hypergeometric distribution. Given two parallel tests  $X$  and  $Y$ , one can check empirically whether or not (23.6.21) holds. In Section 21.6, a similar result was obtained for the Poisson model.

Keats and Lord have given a bivariate negative hypergeometric distribution for graduating scatterplots between parallel test forms, or forms that would be parallel if they had the same number of items (Keats and Lord, 1962, Eq. 2). It can be shown that under this model, the regression of one test on a parallel test is linear. For a generalization of  $h(x)$ , see Keats (1964).

### 23.7 The Beta Distribution of True Scores

Under the binomial error model, whenever the true-score distribution is a two-parameter beta distribution, then the observed-score distribution is negative hypergeometric. The two-parameter beta distribution can be written

$$g(\xi) = \frac{\xi^{a-1}(1-\xi)^{b-n}}{B(a, b-n+1)}, \quad (23.7.1)$$

where  $a > 0$  and  $b > n - 1$ , and where

$$B(a, b-n+1) \equiv \frac{\Gamma(a)\Gamma(b-n+1)}{\Gamma(a+b-n+1)} \quad (23.7.2)$$

is the usual beta function. By carrying out the integration, one can readily verify that

$$\int_0^1 \frac{\xi^{a-1}(1-\xi)^{b-n}}{B(a, b-n+1)} \binom{n}{x} \xi^x (1-\xi)^{n-x} d\xi \equiv h(x), \quad x = 0, 1, \dots, n, \quad (23.7.3)$$

where  $h(x)$  is the negative hypergeometric distribution of (23.6.10). This proves the first statement of this section.

If the observed-score distribution is negative hypergeometric, the true-score distribution is either the two-parameter beta distribution (23.7.1), or some other distribution having identical moments up through order  $n$  (see next section). In either case, the regression of true score on observed score is

given by the linear equation

$$\mathcal{E}(\xi | x) = \alpha_{21}x/n + (1 - \alpha_{21})\mu_X/n, \quad x = 0, 1, \dots, n, \quad (23.7.4)$$

as may be verified from (23.6.1), (23.6.5), (23.6.6), and (23.6.13).

### 23.8 Moments of the True-Score Distribution

Moments of the true-score distribution may be readily expressed in terms of moments of the observed-score distribution by means of (23.4.1), without assuming linear regression or making any other assumptions beyond (23.4.1) itself.

For the observed-score distribution, it is convenient to work with factorial moments. The  $r$ th factorial moment of  $\varphi(x)$  is defined as

$$M_{[r]} \equiv \sum_{x=0}^n x^{[r]} \varphi(x), \quad (23.8.1)$$

where

$$x^{[r]} \equiv x(x - 1) \cdots (x - r + 1) \equiv x!/(x - r)!. \quad (23.8.2)$$

One may readily see that a factorial moment of order  $r$  is merely a linear function of ordinary moments up through order  $r$ . Formulas for computing factorial moments from ordinary moments are given by Kendall and Stuart (1958, Eq. 3.25).

Multiply (23.4.1) by  $x^{[r]}$ ,  $r \leq n$ , and sum on  $x$  to obtain

$$M_{[r]} = \int_0^1 g(\xi) \sum_{x=0}^n x^{[r]} \binom{n}{x} \xi^x (1 - \xi)^{n-x} d\xi.$$

The sum on the right is the  $r$ th factorial moment of the binomial distribution, and this is simply  $n^{[r]} \xi^r$  (Kendall and Stuart, 1958, Eq. 5.11). Substituting this in the foregoing equation, we obtain the result

$$M_{[r]} = n^{[r]} \int_0^1 \xi^r g(\xi) d\xi = n^{[r]} \mu'_r, \quad (23.8.3)$$

where  $\mu'_r$  is the  $r$ th ordinary moment (about the origin) of the true-score distribution. For  $r \leq n$ , (23.8.3) can be divided by  $n^{[r]}$ . Thus if (23.4.1) holds, the  $r$ th moment of the true-score distribution,  $r \leq n$ , can be obtained from the observed-score distribution by the formula

$$\mu'_r = M_{[r]}/n^{[r]}, \quad r = 1, 2, \dots, n. \quad (23.8.4)$$

Another important conclusion is clear from the foregoing: Any  $g(\xi)$  with first  $n$  moments satisfying (23.8.4) will itself satisfy the basic equation (23.4.1). Thus under the binomial error model, the true-score distribution cannot be uniquely

determined from the observed-score distribution. However, the first  $n$  moments of the true-score distribution are determined. As we have already pointed out in Section 10.8, if two distributions have the same first  $n$  moments, they have the same best-fitting (least-squares) polynomial of degree  $n$ . For  $n \geq 25$ , as is usual for mental tests, the two distributions will thus be very much alike, provided that they are both reasonably smooth, without oscillations. As we also noted in Section 10.8, even a few moments of a distribution can often be used to estimate its shape remarkably well.

Next, consider the problem of determining the bivariate moments of the joint distribution of true score and observed score. It is this bivariate distribution that completely describes the properties of the test as a measuring instrument. By (23.2.5), the bivariate distribution of observed score and true score is

$$g(\xi) \binom{n}{x} \xi^x (1 - \xi)^{n-x}.$$

Consequently, as in (23.8.3),

$$\begin{aligned} \mathcal{E}(x^{[r]} \xi^s) &= \int_0^1 \xi^s g(\xi) \sum_{x=0}^n x^{[r]} \binom{n}{x} \xi^x (1 - \xi)^{n-x} d\xi = \int_0^1 \xi^s g(\xi) n^{[r]} \xi^r d\xi \\ &= n^{[r]} \mu'_{r+s}. \end{aligned} \quad (23.8.5)$$

Substitute (23.8.4) into this to find

$$\mathcal{E}(x^{[r]} \xi^s) = \frac{n^{[r]} M_{[r+s]}}{n^{[r+s]}} = \frac{M_{[r+s]}}{(n - r)^{[s]}}, \quad r + s = 1, 2, \dots, n. \quad (23.8.6)$$

Thus for  $r + s \leq n$ , the  $(r, s)$ th bivariate moment of true score and observed score can be obtained directly from the univariate moments of the observed-score distribution.

Since

$$M_{[2]} = \mathcal{E}(X^2) - \mathcal{E}(X)^2, \quad (23.8.7)$$

one quickly finds from (23.8.4) that under the binomial model, the variance of the proportion-correct true score is

$$\sigma_{\xi}^2 = \frac{1}{n^{[2]}} \left[ \sigma_X^2 - \frac{1}{n} \mu_X (n - \mu_X) \right]. \quad (23.8.8)$$

The variance of the number-correct true score ( $\xi \equiv n\xi$ ) is then

$$\sigma_{\xi}^2 = n^2 \sigma_{\xi}^2 = \frac{1}{n-1} [n \sigma_X^2 - \mu_X (n - \mu_X)]. \quad (23.8.9)$$

Under the binomial error model, the ratio of number-correct true-score variance

to observed-score variance is

$$\frac{\sigma_{\xi}^2}{\sigma_X^2} = \frac{n}{n-1} \left[ 1 - \frac{\mu_X(n-\mu_X)}{n\sigma_X^2} \right] = \alpha_{21}, \quad (23.8.10)$$

the Kuder-Richardson formula 21 reliability coefficient.

In classical test theory (Section 3.3), the ratio  $\sigma_{\xi}^2/\sigma_X^2$  is equal to the squared correlation between observed score and true score. The same result,

$$\rho_{X\xi}^2 = \rho_{X\xi}^2 = \sigma_{\xi}^2/\sigma_X^2, \quad (23.8.11)$$

can be proved directly from (23.8.4) and (23.8.6). Thus, by (23.8.10), *the correlation between observed score and true score under the binomial error model is equal to the square root of the Kuder-Richardson formula 21 reliability:*

$$\rho_{X\xi} = \rho_{X\xi} = \sqrt{\alpha_{21}}. \quad (23.8.12)$$

The significance of this result is discussed below.

### 23.9 Relation of the Binomial Error Model to Other Test Theory Models

For dichotomous items, Tucker's formula (1949, Eq. 27) shows the relation of  $\alpha_{21}$  to  $\alpha_{20}$ :

$$\alpha_{20} = \alpha_{21} + \frac{n}{n-1} \frac{n\sigma_{\pi}^2}{\sigma_X^2}, \quad (23.9.1)$$

where

$$\sigma_{\pi}^2 = \frac{1}{n} \sum_{i=1}^n \left( \pi_i - \frac{\mu_X}{n} \right)^2$$

is the variance over items of the item difficulties. Thus  $\alpha_{21}$  is a lower bound to coefficient  $\alpha_{20}$ , which itself is a lower bound to the test reliability. The two coefficients are equal only when all items are of equal difficulty. Figure 23.9.1 shows the relation between estimated values of  $\alpha_{21}$  and  $\alpha_{20}$  for 58 published tests.

Since the correlation between observed score and true score in classical test theory is defined as the square root of the test reliability, and since the binomial error model requires that this correlation equal  $\sqrt{\alpha_{21}}$ , it is clear that either

- 1) the true score of the binomial error model is not the same as the true score of the classical model, or
- 2) the binomial model must be viewed as an oversimplified approximation to the classical model, except when the test items are all of equal difficulty.

In Section 11.9 it was clear that under the item sampling model, the test score of a given examinee is a random variable having a binomial distribution

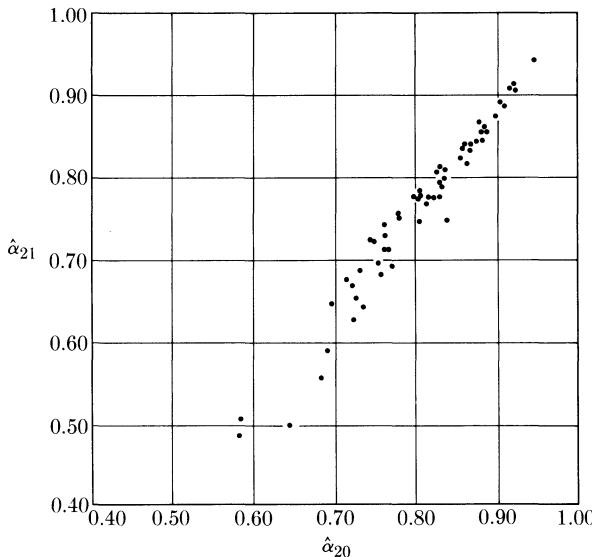


FIG. 23.9.1. Relation between estimates of  $\alpha_{20}$  and formula 20 reliability coefficients for 58 tests. [From F. M. Lord, Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement*, 1959, 19, 233-239. Used by permission.]

(23.2.5) with parameter  $\xi$  representing his *generic* true score. Does this fact justify the binomial model represented by (23.4.1)?

The item sampling model provides good justification for the binomial conditional distribution (23.2.5) for any single examinee. For the basic equation (23.4.1) to hold, however, it is necessary in addition that the observed scores of different examinees be distributed independently of one another. This independence exists whenever each examinee takes a different random sample of  $n$  items. However, independence does not exist in the usual case where all examinees take the same random sample of  $n$  items, as we have pointed out in connection with (8.4.2). Thus the item sampling model does not provide a complete justification for the binomial model of (23.4.1).

### 23.10 The Compound Binomial Error Model

For many purposes, the binomial error model can be viewed as a useful approximation. Is there a better model that could be used? A consideration of latent-trait theory equations based on item characteristic curves leads to a clearer view of this problem.

Consider the case where the test items depend on just one latent trait  $\theta$ . Since, as shown by Theorem 16.13.1, the true score  $\xi$  is functionally related

to  $\theta$  by the equation

$$\xi \equiv \frac{1}{n} \sum_{g=1}^n P_g(\theta), \quad (23.10.1)$$

the conditional distribution of observed score for fixed  $\xi$  is the same as for fixed  $\theta$ . By (16.12.2), this conditional distribution  $h(x | \xi)$  is the compound binomial distribution generated by the probability generating function

$$\prod_{g=1}^n (Q_g + P_g t), \quad (23.10.2)$$

where  $P_g$  and  $Q_g \equiv 1 - P_g$  may be viewed as functions of  $\xi$ .

This line of reasoning clearly leads to a compound binomial error model exactly like the one represented by (23.4.1), except that the binomial "kernel" is replaced by a compound binomial one. To make further progress, however, it is necessary to know more about  $P_g$  as a function of  $\xi$ . If some two-parameter functional form is assumed, such as the normal ogive or logistic, then there are  $2n$  parameters to be estimated before the model can be used in practical work. Even if these  $2n$  parameters are estimated, the sampling errors in the  $2n$  estimates may have a cumulative effect that destroys the utility of the model.

Since approximations appear inevitable, let us try to find some that simplify the problem. Letting

$$\xi = \frac{1}{n} \sum_{g=1}^n P_g,$$

as in (23.10.1), we can expand the compound binomial in powers of  $(P_1 - \xi)$ ,  $(P_2 - \xi), \dots, (P_n - \xi)$ , (see Walsh, 1963). The compound binomial distributions can therefore be written

$$\begin{aligned} \text{Prob}(X = x) &= p_n(x) + \frac{1}{2}nV_2C_2(x) + \frac{1}{3}nV_3C_3(x) + (\frac{1}{4}nV_4 - \frac{1}{8}n^2V_2^2)C_4(x) \\ &\quad + (\frac{1}{5}nV_5 - \frac{5}{6}n^2V_2V_3)C_5(x) + \dots, \quad x = 0, 1, \dots, n, \end{aligned} \quad (23.10.3)$$

where

$$p_n(x) = \begin{cases} \binom{n}{x} \xi^x (1 - \xi)^{n-x} & \text{for } x = 0, 1, \dots, n, \\ 0 & \text{otherwise,} \end{cases} \quad (23.10.4)$$

$$C_r(x) = \sum_{v=0}^r (-1)^{v+1} \binom{r}{v} p_{n-r}(x - v), \quad (23.10.5)$$

$$V_r = \frac{1}{n} \sum_{g=1}^n (P_g - \xi)^r, \quad r = 2, 3, \dots, n. \quad (23.10.6)$$

Series (23.10.3) is a finite series; when all  $n$  terms are used, it is an exact identity. Since the higher-order terms become small, it is practical to truncate the series. A convenient result holds: No matter where (23.10.3) is truncated,

the truncated series is a frequency distribution, provided that the parameters of the series are such that the truncated sum is nonnegative for each value of  $x = 0, 1, \dots, n$ .

If we use only the first term of the series, then we have the binomial model of (23.4.1). The second term involves  $V_2$ , which is an unknown function of  $\xi$ . It is obvious from (23.10.1) and (23.10.6) that  $V_2$  must be zero when  $\xi = 0$  and when  $\xi = 1$ . Some experience with the normal ogive model suggests that a fair approximation to  $V_2$  is given by the expression

$$V_2 \doteq 2k\xi(1 - \xi)/n, \quad (23.10.7)$$

where  $k$  is a parameter to be determined. This approximation provides a working mathematical model for true-score theory that can be tested out on real data and then modified, if desired, to improve predictions of observable results (see Section 23.11).

We are now in a position to choose the parameter  $k$  to reflect the characteristics of the particular test under consideration. For example, by means of an item analysis, we can accurately estimate the variance of item difficulties  $\sigma_\pi^2$ . With the aid of (23.9.1), a formula is then available (Lord, 1965) for obtaining  $k$  so that under the resulting model, *the correlation between observed score and true score will always equal  $\sqrt{\alpha_{20}}$* :

$$k = \frac{n^2(n - 1)\sigma_\pi^2}{2[\mu_X(n - \mu_X) - \sigma_X^2 - n\sigma_\pi^2]}. \quad (23.10.8)$$

This is much more nearly satisfactory than was the correlation equal to  $\sqrt{\alpha_{21}}$ , obtained under the binomial model. The superiority of the compound binomial model is particularly evident when the model is used to predict the bivariate distribution of scores on parallel test forms, to be discussed in the following section. We should expect this, since  $\alpha_{20}$  is a better approximation to the correlation between such forms than is  $\alpha_{21}$ .

Under the model based on (23.10.8), the squared correlation  $\rho_{X\xi}^2$  between observed score and true score varies with  $n$  according to the Spearman-Brown formula. Similarly, also, the variance of the observed score varies with  $n$  as in Eq. (4.3.10); that is, if test  $X'$  is  $L$  times as long as test  $X$ , then

$$\sigma_{X'}^2 = L\sigma_X^2[1 + (L - 1)\rho_{X\xi}^2].$$

### 23.11 Solving for the Distribution of True Scores

The basic equation under consideration here is

$$\varphi(x) = \int_0^1 g(\xi) h(x | \xi) d\xi, \quad x = 0, 1, \dots, n, \quad (23.11.1)$$

where  $h(x | \xi)$  is some approximation to the compound binomial generated by (23.10.2). Consider the case where  $h(x | \xi)$  is known and  $\varphi(x)$  is estimated from

an actual distribution of observed scores. The problem is to make inferences about the nature of  $g(\xi)$ .

For simple cases, an approach parallel to that of Section 23.8 allows us to express the first  $n$  moments of  $g(\xi)$  in terms of the first  $n$  moments of  $\varphi(x)$ . Beyond this, however, it is clear from Section 23.7 that  $g(\xi)$  is not uniquely determined by (23.11.1). On the other hand, as we have pointed out in connection with Eq. (23.8.4), if two different  $g(\xi)$  having the same first  $n$  moments are both smooth in some sense, then they cannot differ much from each other. Consequently it would be tolerably satisfactory to have a good estimation method that produces a "smooth" estimated  $g(\xi)$ .

We can obtain an analytic expression for  $g(\xi)$  that satisfies (23.11.1) by applying standard methods for solving integral equations to (23.11.1). (Such standard methods can be found in Tricomi, 1957, for example.) These methods could produce good results if  $\varphi(x)$  were known for the population of examinees, but they usually produce absurd results when  $\varphi(x)$  is replaced by the observed distribution for a sample of examinees. Tikhonov (1963a, 1963b), Twomey (1965), and others have discussed available nonstatistical methods for working from fallible data; Lord (1967), Maritz (1966), and Robbins (1964) have discussed related statistical inference problems. However, much work still remains to be done on this problem from the point of view of statistical inference.

If adequate estimates of or approximations to true-score distributions can be obtained by such procedures, then many practical results may be obtained. Most of these flow from the ability of the model to predict the bivariate distribution of two homogeneous measures of the same psychological trait, using only the information in the univariate (marginal) distributions.

Suppose two tests  $X$  and  $X'$  have been administered to the same population, and suppose that their respective true-score distributions  $g(\xi)$  and  $g'(\xi')$  are independently estimated from the univariate distributions of  $X$  and  $X'$ , respectively. If the two tests "measure the same psychological trait", we take this to mean that  $\xi$  and  $\xi'$  differ only by choice of metric; that is, they are related by a monotonic transformation,  $\xi' = \psi(\xi)$ , say. The two tests may differ in length, difficulty, etc. The key point is that the transformation  $\psi$  can be determined from estimates of  $g$  and  $g'$ , because for all values of  $\xi$  it must satisfy

$$\int_0^\xi g(\xi) d\xi \equiv \int_0^{\psi(\xi)} g'(\xi') d\xi'. \quad (23.11.2)$$

When  $g$  and  $g'$  have been estimated for various values of  $\xi$ , this equation is used to estimate the corresponding values of  $\psi(\xi)$ .

If the two tests are homogeneous, then an equation like (23.11.1) is assumed to hold for their bivariate distribution  $\varphi(x, x')$ :

$$\varphi(x, x') = \int_0^1 g(\xi) h(x | \xi) h(x' | \xi') d\xi, \quad x = 0, 1, \dots, n; x' = 0, 1, \dots, n'. \quad (23.11.3)$$

Since the relationship  $\zeta' \equiv \psi(\zeta)$  can be estimated,  $h(x' | \zeta')$  can be treated as a known function of  $\zeta$ , as can the other terms in the integrand. The integration in (23.11.3) can then be carried out by numerical methods to produce an estimate of  $\varphi(x, x')$ . Thus *for two tests measuring the same trait, the bivariate distribution of observed scores  $\varphi(x, x')$  can be estimated from the univariate distribution of observed scores on each test.*

Furthermore, once  $\psi(\zeta)$  has been determined, this function remains invariant regardless of the group tested so long as tests  $X$  and  $X'$  still have true scores  $\zeta$  and  $\zeta'$ . Thus, once  $\psi(\zeta)$  has been estimated, the bivariate distribution of  $X$  and  $X'$  can be estimated for various groups that have taken only one of the two tests. This can be a useful fact in preparing test norms, since the frequency distribution of a long test for a large normative group can be inferred by administering a short form to the large group instead of the long form.

This same general approach can be used to estimate the effect of selecting examinees on an unreliable score  $X$  instead of on the measure of real interest  $\zeta$ . For example, suppose a college admits only applicants who score one standard deviation above the national mean on a rather unreliable measure of scholastic aptitude. What is the distribution of true scores for the admitted students? Is it positively or negatively skewed? How many admitted students will have a true aptitude below the national mean?

The method also can be used to investigate the question, whether two tests  $X$  and  $X'$  really measure the same psychological trait. The psychometrician currently has no adequate way of answering this important question! Since methods for estimating  $g(\zeta)$  are still only experimental, we shall not give a more detailed treatment here. Some empirical tryouts have been reported by Lord (1965, 1967).

### References and Selected Readings

- BLYTH, C. R., and D. W. HUTCHINSON, Table of Neyman-shortest unbiased confidence intervals for the binomial parameter. *Biometrika*, 1960, **47**, 381–392.
- CROW, E. L., Confidence intervals for a proportion. *Biometrika*, 1956, **43**, 423–435.
- ERDELYI, A. (Ed.), *Higher transcendental functions*, Volume I. New York: McGraw-Hill, 1953.
- KEATS, J. A., Some generalizations of a theoretical distribution of mental test scores. *Psychometrika*, 1964, **29**, 215–231.
- KEATS, J. A., and F. M. LORD, A theoretical distribution for mental test scores. *Psychometrika*, 1962, **27**, 59–72.
- KENDALL, M. G., and A. STUART, *The advanced theory of statistics*. Vol. 1: *Distribution theory*. New York: Hafner, 1958.
- LORD, F. M., A strong true-score theory, with applications. *Psychometrika*, 1965, **30**, 239–270.

- LORD, F. M., Estimating true-score distributions in psychological testing—an empirical Bayes estimation problem. *Research Bulletin 67-87*. Princeton, N.J.: Educational Testing Service, 1967.
- MARITZ, J. S., Smooth empirical Bayes estimation for one-parameter discrete distributions. *Biometrika*, 1966, **53**, 417-429.
- ROBBINS, H., The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics*, 1964, **35**, 1-20.
- TIKHONOV, A. N., Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 1963, **4**, 1035-1038. (a)
- TIKHONOV, A. N., Regularization of incorrectly posed problems. *Soviet Mathematics Doklady*, 1963, **4**, 1624-1627. (b)
- TRICOMI, F. G., *Integral equations*. New York: Interscience, 1957.
- TUCKER, L. R., A note on the estimation of test reliability by the Kuder-Richardson formula (20). *Psychometrika*, 1949, **14**, 117-120.
- TWOMEY, S., The application of numerical filtering to the solution of integral equations encountered in indirect sensing measurements. *Journal of the Franklin Institute*, 1965, **279**, 95-109.
- WALSH, J. E., Corrections to two papers concerned with binomial events. *Sankhyā*, Series A, 1963, **25**, 427.

# TRUE SCORES, FACTORS, AND LATENT TRAITS

## 24.1 Introduction

The underlying concept, central to almost every model studied in this book, is that of the latent trait. We shall not attempt to present a complete development of latent trait theory, since this would be beyond our needs. However, we shall develop the essential features of the general theory, and attempt to show the interrelationship of many of the concepts of latent trait theory in a manner that suggests how a more complete development might proceed. We shall also catalogue some mathematical and statistical techniques and results which researchers may find useful in future investigations. Much of the material presented here has appeared in less integrated treatments earlier in the text.

In Section 24.2, we present an axiomatic treatment of the multiple factor analytic model. The conceptual value of the model is substantial, and when it is presented in a manner parallel to the development in Chapter 2, one can clearly see that the model has great generality. The factor analytic model, however, is not a latent trait model in one technical sense, which we shall introduce in this chapter. In Section 3, we show how the factor analytic and the true-score models provide distinct decompositions of the observed-score random variable into overlapping latent variables. In Section 4, we suggest one definition of latent trait models, namely, as those models that adopt an assumption of local independence. In Section 5, we characterize a strong assumption of local independence and show its relationship to the assumption of experimental independence. For most practical work, so strong an assumption is not needed. In Section 6, we present a general relationship between true scores and latent traits. Finally, in Section 7, we present a very general statement of latent trait models and derive its important special cases, some of which are not discussed elsewhere in this book.

## 24.2 The Multiple Factor Analytic Model

The classical test theory model deals with two kinds of random variables, *manifest* or observable variables and *latent* or unobservable variables. The true- and error-score random variables of the classical model are examples of latent variables, as are the Platonic true scores of the pullet-cockerel example (Sec-

tion 2.9). The definition of true and error scores in the classical model permits the decomposition of the observed score into two orthogonal latent components, the first associated with variation between persons and the second with residual or within-person variation. Thus we say that the classical true-score model is an example of a latent variable model. When multiple measurements are made on each of a number of persons, another kind of latent variable model is useful for analyzing and decomposing the observed variables into latent components so as to "explain" the correlation structure of the observed variables.

Let  $\mathbf{x}' \equiv (x_1, x_2, \dots, x_i, \dots, x_n)$  be a row vector of  $n$  observed-score random variables corresponding to mental measurements or other variables (age, for example) and having finite variances. Let  $\boldsymbol{\mu}'$  be the row vector whose elements are the expected values of elements of  $\mathbf{x}'$ ,  $\mathcal{E}\mathbf{x}' \equiv \boldsymbol{\mu}'$ . It is our desire to explain the correlational structure of the observed variables by expressing them as weighted linear combinations of a *lesser* number of latent variables with a residual component for each observed variable. Let  $\boldsymbol{\theta}' \equiv (\theta_1, \theta_2, \dots, \theta_j, \dots, \theta_k)$  be a row vector of  $k$  latent variables known as *common factors*, with  $k < n$ . Both  $\mathbf{x}$  and  $\boldsymbol{\theta}$  are defined in some fixed population of persons.

Let  $\mathbf{\Lambda} \equiv \|\lambda_{ij}\|$  be an  $n \times k$  matrix of constants (specifically, the weights relating  $\mathbf{x}$  to  $\boldsymbol{\theta}$ ) called *factor loadings*. Also, let us require that these constants are such that each column of  $\mathbf{\Lambda}$  has two or more nonzero elements; this requirement identifies each element of  $\boldsymbol{\theta}$  as a common factor, since it implies that each such element is common to at least two of the manifest variables. We then define the  $n$ -element vector random variable  $\mathbf{u}$  by the relation

$$\text{i) } \mathbf{u} \equiv (\mathbf{x} - \boldsymbol{\mu}) - \mathbf{\Lambda}\boldsymbol{\theta}. \quad (24.2.1)$$

The elements of  $\mathbf{u}$  are called the *unique factors* corresponding to the elements of  $\mathbf{x}$ . These are the residuals when the common parts (and the grand means) have been subtracted from  $\mathbf{x}$ . Rewriting (24.2.1), we have the basic equation of the *linear factor analytic model*:

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Lambda}\boldsymbol{\theta} + \mathbf{u}. \quad (24.2.2)$$

Let  $\mathbf{D}_\omega^2$  be the dispersion matrix of  $\mathbf{u}$ , the diagonal elements  $\omega_{ii} \equiv \omega_i^2$  being the variances of the elements of  $\mathbf{u}$ . We now assume that

$$\text{ii) } \mathbf{D}_\omega^2 \text{ is diagonal.} \quad (24.2.3)$$

Assumption (24.2.3) states that the components of  $\mathbf{u}$  are uncorrelated with each other. This implies that  $\boldsymbol{\theta}$  accounts for all of the correlations among the elements of  $\mathbf{x}$ ; that is, it implies that the partial correlation between any two elements of  $\mathbf{x}$  is zero when  $\boldsymbol{\theta}$  is held constant. Since measurements typically contain error, we shall ordinarily have  $\omega_i^2 > 0$  for all  $i$ . This implies that none of the elements of  $\mathbf{x}$  is completely determined by a linear combination of the elements of  $\boldsymbol{\theta}$ ; that is, none of the unique factors has a zero variance.

We also assume that

$$\text{iii) } \mathcal{E}(\mathbf{u} | \boldsymbol{\theta}) = \mathbf{0} \quad \text{for all } \boldsymbol{\theta}: \quad (24.2.4)$$

For any specification of each of the elements of  $\boldsymbol{\theta}$ , the expected value of  $\mathbf{u}$  is  $\mathbf{0}$ . Assumption (24.2.4) is equivalent to an assumption of linear regression of  $\mathbf{x}$  on  $\boldsymbol{\theta}$ , namely,

$$\mathcal{E}(\mathbf{x} | \boldsymbol{\theta}) = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\theta}. \quad (24.2.5)$$

We prove this by applying (24.2.4) to (24.2.1), obtaining

$$\begin{aligned} \mathcal{E}(\mathbf{u} | \boldsymbol{\theta}) &= \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu}) - \boldsymbol{\Lambda}\boldsymbol{\theta} | \boldsymbol{\theta}] \\ &= \mathcal{E}(\mathbf{x} | \boldsymbol{\theta}) - \boldsymbol{\mu} - \boldsymbol{\Lambda}\boldsymbol{\theta} = \mathbf{0}, \end{aligned}$$

or

$$\mathcal{E}(\mathbf{x} | \boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\mu}.$$

Assumption (24.2.4) also implies that each of the elements of  $\boldsymbol{\theta}$  is uncorrelated with each of the elements of  $\mathbf{u}$ , and that  $\mathcal{E}(\mathbf{u}) = \mathbf{0}$ . Thus the variables  $\mathbf{u}$  have the orthogonality and zero expectation properties of the usual errors of the classical test-theory model. In general, however, they are not identical with the classical error variables, as we shall demonstrate later.

We denote the dispersion (variance-covariance) matrix of  $\mathbf{x}$  by  $\boldsymbol{\Sigma} \equiv \|\sigma_{ii'}\|$ , the dispersion matrix of  $\boldsymbol{\theta}$  by  $\boldsymbol{\Psi} \equiv \|\psi_{jj'}\|$ , and the covariances between the vector variables  $\mathbf{x}$  and  $\boldsymbol{\theta}$  by the  $n \times k$  matrix  $\boldsymbol{\Gamma} = \|\gamma_{ij'}\|$ . The dispersion matrix of  $(\mathbf{x}', \boldsymbol{\theta}')$  is then

$$\begin{vmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}' & \boldsymbol{\Psi} \end{vmatrix}.$$

An additional simplifying assumption is that

$$\text{iv) } \boldsymbol{\Psi} \text{ is nonsingular.} \quad (24.2.6)$$

Assumption (24.2.6) asserts that the  $k$  common factors form a linearly independent set, which is to say that the model does not hold for some smaller value of  $k$ . With no loss of generality, we may adjust  $\boldsymbol{\Lambda}$  whenever desirable so that

$$\text{v) } \psi_{jj} = 1 \quad \text{for all } j: \quad (24.2.7)$$

The common factors have unit variance. The result

$$\mathcal{E}(\boldsymbol{\theta}) = \mathbf{0} \quad (24.2.8)$$

follows from (24.2.2), (24.2.4), and the definition  $\mathcal{E}(\mathbf{x}) = \boldsymbol{\mu}$ .

A basic equation of the factor analytic model involves a decomposition of  $\boldsymbol{\Sigma}$ , the variance-covariance matrix of  $\mathbf{x}$ , into two parts. We state this equation as Theorem 24.2.1.

**Theorem 24.2.1**

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Psi}\mathbf{A}' + \mathbf{D}_\omega^2. \quad (24.2.9)$$

*Proof*

$$\begin{aligned} \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] &= \mathcal{E}[(\mathbf{A}\boldsymbol{\theta} + \mathbf{u})(\mathbf{A}\boldsymbol{\theta} + \mathbf{u})'] \\ &= \mathcal{E}(\mathbf{A}\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{A}' + \mathbf{A}\boldsymbol{\theta}\mathbf{u}' + \mathbf{u}\boldsymbol{\theta}'\mathbf{A}' + \mathbf{u}\mathbf{u}') \\ &= \mathbf{A}\mathcal{E}(\boldsymbol{\theta}\boldsymbol{\theta}')\mathbf{A}' + \mathcal{E}(\mathbf{u}\mathbf{u}') \\ &= \mathbf{A}\boldsymbol{\Psi}\mathbf{A}' + \mathbf{D}_\omega^2. \quad \square \end{aligned}$$

In Chapter 2, we saw that a single observed-score random variable  $X$  could always be decomposed into the sum of two orthogonal random variables  $T$  and  $E$ , the true- and error-score random variables. It is no more difficult to show that for any  $n$ -element vector  $\mathbf{x}$  of observed-score random variables, the factor analytic decomposition into  $k$  common (latent) variables and  $n$  unique (latent) variables is always possible, provided that  $k = n - 1$ . Hence it is always possible to find an  $(n - 1)$ -element vector  $\boldsymbol{\theta}$  such that  $\boldsymbol{\Sigma} - \mathbf{A}\boldsymbol{\Psi}\mathbf{A}'$  is a diagonal matrix  $\mathbf{D}_\omega^2$ . The generality of this decomposition has been described by Guttman (1957). For  $k = n - 1$ , then, the factor analytic model is not a model at all, but rather only a tautology.

If this factorization violates the assumption (24.2.6), then a factorization that satisfies this assumption must exist for some smaller value of  $k$ . That is to say, that for  $k = n - 1$ , the assumptions of the factor-analytic model are always satisfied, and the model places no restrictions on the kinds of data that it will accurately model. However, the assumption that the factor analytic decomposition can be obtained for some value  $k < n - 1$  is an assumption that may be false. When the model does hold, we shall say that the matrix  $\boldsymbol{\Sigma}$  has been factored into the matrices  $\mathbf{A}$ ,  $\boldsymbol{\Psi}$ , and  $\mathbf{D}_\omega^2$ . In application, the factor analytic model is useful primarily when a representation can be obtained, exactly or approximately, for  $k$  very much less than  $n$ , for then we may say that the intercorrelations of  $\mathbf{x}$  have been explained in terms of a smaller number of hypothetical factors. This explanation is obtained by giving psychologically meaningful names to the various factors that are constructed. The relative magnitudes of the elements of any row of  $\mathbf{A}$ , the vector of factor loadings, determine the relative extent to which that test measures each of the latent variables.

Since  $\mathbf{D}_\omega^2$  is a diagonal matrix, it is clear from (24.2.8) that the nondiagonal elements of  $\mathbf{A}\boldsymbol{\Psi}\mathbf{A}'$  are the same as those of  $\boldsymbol{\Sigma}$  and that the diagonal elements of  $\mathbf{A}\boldsymbol{\Psi}\mathbf{A}'$  are equal to those of  $\boldsymbol{\Sigma}$  minus those of  $\mathbf{D}_\omega^2$ . Let  $\mathbf{D}_\sigma^2 = \text{diag } \boldsymbol{\Sigma}$  be a diagonal matrix whose diagonal elements are equal to those of  $\boldsymbol{\Sigma}$ . Then the diagonal elements of  $\mathbf{D}_\sigma^{-1}\mathbf{A}\boldsymbol{\Psi}\mathbf{A}'\mathbf{D}_\sigma^{-1}$ , the standardized form of  $\mathbf{A}\boldsymbol{\Psi}\mathbf{A}'$ , may be denoted by  $\rho_i^2$ . These values will be called the *communalities* of the observed variables  $\mathbf{x}$ . *The  $i$ th communality is the squared multiple correlation between the  $i$ th observed variable and the set of  $k$  factors.*

A factor analytic decomposition for any value of  $k$ , however, is never unique. For suppose that  $(\boldsymbol{\theta}, \mathbf{\Lambda}, \mathbf{\Psi})$  provides a factor analytic decomposition for some value of  $k$ . By this we mean that we can determine a vector  $\boldsymbol{\theta}$  of common factors, so that the model holds, where  $\mathbf{\Lambda}$  is the matrix of factor loadings and  $\mathbf{\Psi}$  is the dispersion matrix of  $\boldsymbol{\theta}$ . Let  $\mathbf{T}$  be an arbitrary nonsingular matrix of order  $k$ , and let

$$\mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{T}^{-1}, \quad (24.2.10)$$

$$\mathbf{\Psi}^* = \mathbf{T}\mathbf{\Psi}\mathbf{T}', \quad (24.2.11)$$

$$\boldsymbol{\theta}^* = \mathbf{T}\boldsymbol{\theta}. \quad (24.2.12)$$

Then  $\mathbf{\Lambda}^*\boldsymbol{\theta}^* = \mathbf{\Lambda}\boldsymbol{\theta}$ , and hence

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Lambda}^*\boldsymbol{\theta}^* + \mathbf{u};$$

and  $\mathbf{\Psi}^*$  is the dispersion matrix of  $\boldsymbol{\theta}^*$ . Thus  $(\boldsymbol{\theta}^*, \mathbf{\Lambda}^*, \mathbf{\Psi}^*)$  provides a factor analytic decomposition of  $\mathbf{x}$  distinct from that provided by  $(\boldsymbol{\theta}, \mathbf{\Lambda}, \mathbf{\Psi})$ . In general, therefore, if a factor analytic decomposition exists for some value of  $k$ , then there must be an infinite number of solutions for that  $k$ . The problem of which decomposition to choose is called the *problem of rotation*.

To limit the number of solutions that may be possible for any value of  $k$  and to finally identify a unique decomposition, we may place various restrictions on the model, in the form of conditions that the latent variables  $\boldsymbol{\theta}$  must satisfy. To begin with, we can limit the number of solutions somewhat by requiring that the common factors  $\boldsymbol{\theta}$  be *orthogonal*, that is, that  $\mathbf{\Psi}$  be a diagonal matrix. If the  $\boldsymbol{\theta}$  are not orthogonal, then they are said to be *oblique*. If an oblique factorization exists, then there will always exist a transformation that yields an orthogonal factorization when used with Eqs. (24.2.10) through (24.2.12). The restriction to orthogonal factors permits a very convenient mathematical simplification of the model. However, many factor analysts and test theorists (including the present writers) find factor decompositions into oblique factors more meaningful and useful psychologically.

Moreover, even the restriction to orthogonal factors does not provide a unique factor analytic decomposition. For suppose that  $(\boldsymbol{\theta}, \mathbf{\Lambda}, \mathbf{\Psi})$  provides an orthogonal factorization of  $(\mathbf{x}, \boldsymbol{\Sigma})$ , and that  $\mathbf{T}$  is an arbitrary orthogonal matrix. Then (24.2.10) through (24.2.12) define an alternative orthogonal factorization of  $(\mathbf{x}, \boldsymbol{\Sigma})$ . To obtain a unique factorization of a matrix, we must place further restrictions on the model given here. A discussion of the various further restrictions that may be placed on the factor analytic model and the estimation procedures associated with each such restricted model are beyond the scope of this book. The interested reader is referred to the appropriate parts of Thurstone (1947), Harman (1960), and Jöreskog (1962, 1963, 1966).

In practice, an *exact fit* of the model is seldom possible for very small  $k$ , and the researcher must be satisfied with, at most, a *good fit*. If the fit were perfect, then, in the population, the residual covariance matrix  $\boldsymbol{\Sigma} - \mathbf{\Lambda}\mathbf{\Psi}\mathbf{\Lambda}'$  would be diagonal. Factor analytic methods are procedures for selecting  $\mathbf{\Lambda}$  so as to mini-

mize some function of the off-diagonal elements of this residual (uniqueness) covariance matrix in the sample (Anderson, 1959, and Anderson and Rubin, 1956). Browne (1967) has distinguished three main types of factor analysis by the function of the residual matrix to be minimized in each case. Type I involves a minimization of the “within set correlation of residual variates” (see Rozeboom, 1965), a particularly attractive criterion. When a multivariate normal model is assumed, the estimates obtained by the maximum likelihood method (Lawley, 1940) satisfy this minimization criterion *regardless of the true distribution of the data*. Rao (1955) shows that when this model holds, a canonical correlation characterization yields equations for the factor loadings that are equivalent to the maximum likelihood equations. Jöreskog (1966) provides an effective computational procedure for the type I solution, which has the desirable property that the sample residual correlation matrix is always Gramian.

The type II solution identified by Browne is the alpha factor analysis of Kaiser and Caffrey (1965). The type III solution is the standard modified principal components solution introduced by Thomson (1934). Harman and Jones (1966) have recently developed a new computational procedure for minimizing this same criterion. Neither the type II nor type III solutions have the very desirable Gramian residual matrix properties of the type I solution.

### 24.3 A Comparison between the Factor Analytic and True-Score Decompositions

To relate the factor analytic model more closely to the classical test theory model, it is convenient to decompose the uniqueness (latent) random variable. To facilitate this, let us denote the  $\mathbf{u}$  of the previous section by  $\Upsilon\mathbf{u}$ . The required decomposition follows the method of Sections 2.2 through 2.7, and yields the following generalization of (24.2.2):

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\theta} + \Upsilon\mathbf{u} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\theta} + \mathbf{Bs} + \mathbf{C}\tilde{\mathbf{e}}, \quad (24.3.1)$$

where  $\mathbf{B} \equiv \|b_i^2\|$  and  $\mathbf{C} \equiv \|c_i^2\|$  are  $n \times n$  diagonal matrices of constants (weights);  $\mathbf{u} = \{u_i\}$ ,  $\mathbf{s} = \{s_i\}$ , and  $\tilde{\mathbf{e}} = \{\tilde{e}_i\}$  are vector variables; and  $\Upsilon \equiv \|v_i\|$  is an  $n \times n$  diagonal matrix of constants chosen so that the elements of  $\mathbf{u}$ ,  $\mathbf{s}$ , and  $\tilde{\mathbf{e}}$  have unit variance and so that  $\Upsilon\mathbf{u} = \mathbf{Bs} + \mathbf{C}\tilde{\mathbf{e}}$ . The latent variables  $\mathbf{s}$  are called the *specific factors*; and the  $\mathbf{e} = \mathbf{C}\tilde{\mathbf{e}}$  are the usual *error random variables* and the  $\tau = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\theta} + \mathbf{Bs}$  the usual true-score random variables of the classical test theory model. Schematically,

$$\mathbf{x} = \underbrace{\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\theta} + \mathbf{Bs} + \mathbf{C}\tilde{\mathbf{e}}}_{\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\theta} + \Upsilon\mathbf{u}} + \mathbf{e} \quad (24.3.2)$$

where the first line gives the test theory decomposition and the third line gives the factor analytic decomposition.

For illustrative purposes, it is convenient to take orthogonal common factors and to assume that each of the random variables  $x$ ,  $z$ ,  $s$ ,  $u$ , and  $\tilde{e}$  has unit variance. Then

$$\rho_i^2 = \sum_{j=1}^k \lambda_{ij}^2. \quad (24.3.3)$$

The  $i$ th communality is the sum of the squares of the elements of the  $i$ th row of the matrix  $\Lambda$  of factor loadings. The total variance, the communality, the unique variance (the uniqueness), the specific variance (the specificity), and the error variance are then given by the following relations:

total variance	$1$	$= \rho_i^2 + b_i^2 + c_i^2 = \rho_i^2 + v_i^2,$
communality	$\rho_i^2$	$= \rho_i^2 = 1 - v_i^2,$
uniqueness	$v_i^2$	$= b_i^2 + c_i^2 = 1 - \rho_i^2,$
specificity	$b_i^2$	$= b_i^2 = v_i^2 - c_i^2,$
error variance	$c_i^2$	$c_i^2 = 1 - \rho_i^2 - b_i^2,$
reliability	$= 1$	$- c_i^2 = \rho_i^2 + b_i^2.$

(24.3.4)

In this special case, the  $i$ th communality is the percent of total variance that the  $i$ th variable shares with the set of common factor variables. If the  $i$ th specificity is zero, then the  $i$ th communality is the reliability of variable  $x_i$ .

An important lesson of test theory is that the vectors of factor loadings cannot be compared with each other to determine the relative extent to which the true components of each test measure the various underlying factors. This could be done only by a factor analysis of true scores. Similarly it is clear that such comparisons of factor analyses of observed score are not independent of test length.

The multiple-factor analytic model stated in this section is a completely general formulation which includes a number of special cases. The simplest special case, and yet one of considerable test theoretic interest, is the *single-factor model*. The multiple-factor model reduces to the single-factor model if one assumes that the multiple-factor model holds for  $k = 1$ , that is, if one assumes that there is a single factor common to each of the measurements.

Spearman (1904, 1927) proposed that all psychological tests measure a single "general intelligence factor" and a unique factor, peculiar to that test only. Modern factor analytic theory has discarded the idea of a single common factor for most situations, finding the multiple-factor approach of Thurstone (1947) to be more accurate and more useful. However, there is one situation in which the concept of a single underlying factor is a reasonable and a useful assumption. This situation occurs whenever relatively similar components are put together to form a composite measurement. The items of a vocabulary test are very similar to each other, but quite dissimilar to the items of an arithmetic test. In such a case, the components are specifically chosen so that they es-

sentially measure the same factor, although they perhaps have some (small) specific variation. Here we would say that the test measures a single underlying factor.

#### 24.4 Factors as Latent Traits

The factor analytic model is one of a number of models that give concrete form to concepts used in theoretical explanations of human behavior in terms of *latent traits*. In any theory of latent traits, one supposes that human behavior can be accounted for, to a substantial degree, by isolating certain consistent and stable human characteristics, or *traits*, and by using a person's values on those traits to predict or explain his performance in relevant situations.

For example, to predict a person's performance in a graduate program in psychometrics, we would be interested to know his value or score on a particular combination of traits but we would not be interested to know his value on other traits. Traits of interest would include quantitative aptitude, verbal aptitude, mathematical ability, subject-matter knowledge in psychology, and also, perhaps, some less well explicated traits such as perseverance and creativity. Other traits, such as finger dexterity and musical aptitude, would be of little or no interest.

The problem of identifying and defining such traits in terms of observable variables and determining what traits are important in a given behavioral context is the major problem in any theory of psychology. In studying any particular behavioral context, the psychologist will bring to bear the relevant theoretical formulation and will perhaps employ one or more models within that theory as a basis for designing experimental studies and for analyzing the data from these studies. It is this interplay of theory, model, and data that adds to scientific knowledge. Though much of psychological theory is based on a trait orientation, it is not necessary that these traits exist in any physical or physiological sense. It is sufficient for most purposes of psychology that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if his value on these traits substantially determined his behavior.

The factor analytic model described in previous sections does not fit the typical test item, for which the observed-score random variable takes values of one or zero, according to whether the subject's response is correct or incorrect. The difficulty arises from the fact that the common factor variables  $\theta$ , and also the unique factor variables, are usually conceived as taking continuous values, and in general, of course, it is a contradiction to consider a discrete-valued random variable to be a linear combination of a set of continuous variables. One way around this difficulty is to consider the discrete observations as resulting from a data reduction on some underlying set of continuous variables; then, under certain assumptions, one may apply factor analysis to these variables, using tetrachoric correlations (see Chapter 15). But generally other kinds of models are more suitable, and we shall introduce these later in the chapter.

At this point, however, it will be convenient to describe and then make explicit the concept of a single underlying trait, and to define and characterize the conditions under which this concept can be properly applied. We might give one possible interpretation of this term by saying that an individual's performance depends on a single underlying trait if, given his value on that trait, nothing further can be learned from him that can contribute to the explanation of his performance. The proposition is that the latent trait is the only important factor and, once a person's value on the trait is determined, the behavior is random, in the sense of statistical independence (Anderson, 1959).

## 24.5 Conditional Independence

The assumption of local (conditional) independence, in its various forms, is the foundation of latent trait theory, and its nature and force become clear only through repeated study of its uses. We have already used this assumption in a number of places, in both its strong and weak forms. We shall now develop a further facet of this concept, by redefining and characterizing it in terms of concepts developed in Chapter 2. With this redefinition and characterization, we shall be able to restate the discussion of the previous section in a very precise way, giving precise conceptual and mathematical meaning to the notion of the single underlying trait.

Within a given population  $\mathcal{P}$ , or any subpopulation (finite or infinite) of  $\mathcal{P}$ , the assumption of local (conditional) independence states that for fixed  $\boldsymbol{\theta}$ , the joint distribution of  $X_{1*}, X_{2*}, \dots, X_{n*}$  factors into the product of the marginal distribution functions; in other words, given  $\boldsymbol{\theta}$ , the variables  $X_{i*}$  are independently distributed. Using the symbol  $F$  generically to indicate a distribution function, we express this as

**Definition 24.5.1.** Measurements  $g, h, \dots$  are *conditionally independent*, if for fixed  $\boldsymbol{\theta}$ , and any subpopulation of  $\mathcal{P}$ ,

$$F(x_1, x_2, \dots, x_n | \boldsymbol{\theta}) = F(x_1 | \boldsymbol{\theta})F(x_2 | \boldsymbol{\theta}) \cdots F(x_n | \boldsymbol{\theta}). \quad (24.5.1a)$$

An alternative and equivalent formula is

$$F(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n; \boldsymbol{\theta}) = F(x_i | \boldsymbol{\theta}) \quad \text{for all } i. \quad (24.5.1b)$$

Concerning local independence, Anderson (1959, p. 11) wrote:

Apart from any mathematical reason for such an assumption there are psychological or substantive reasons. The proposition is that the latent quantities are the only important factors and that once these are determined behavior is random (in the sense of statistical independence). In another terminology, the set of individuals with specified latent characteristics are "homogeneous".

Anderson did not pursue this discussion further. An excellent heuristic discussion of this concept may be found in Lazarsfeld (1958), however. Lazars-

feld's discussion also suggests that the assumption of experimental independence is part of the assumption of local independence.

In the context of the model development in this section, we may formally state the homogeneity condition as

**Definition 24.5.2.** Measurement  $g$  is *conditionally homogeneous* if the conditional error distributions are identical:

$$F_{g,a}(x_{ga} \mid \boldsymbol{\theta}) \equiv F_{g,a'}(x_{ga'} \mid \boldsymbol{\theta}), \quad (24.5.2)$$

where  $a$  and  $a'$  are arbitrary elements of  $\mathcal{P}$ .

It would seem appropriate to refer to this condition as *the assumption of homogeneous errors*, in the sense suggested by Anderson. This assumption states that all persons with the same latent trait values have the same error distributions over replications. We see that the distributions of (24.5.2) are properly called error distributions by noting that the true score  $\tau_{ga}$  is a mathematical function of  $\boldsymbol{\theta}$ . Hence, for randomly selected people, it follows that  $F(X_{g*} \mid \tau_g) = F(X_{g*} \mid \boldsymbol{\theta})$ , and the first of these is just a translation of the conditional error distribution  $F(X_{g*} - T_{g*} \mid \tau_g)$ .

Now, for notational simplicity, we omit the arguments of the function  $F$  and write

$$F_{gg',*} \mid \boldsymbol{\theta} = \mathcal{E}_a F_{gg',a} \mid \boldsymbol{\theta}.$$

By experimental independence,

$$F_{gg',a} = F_{g,a} F_{g',a};$$

hence

$$F_{gg',a} \mid \boldsymbol{\theta} = (F_{g,a} \mid \boldsymbol{\theta})(F_{g',a} \mid \boldsymbol{\theta}).$$

Thus

$$F_{gg',*} \mid \boldsymbol{\theta} = \mathcal{E}_a [(F_{g,a} \mid \boldsymbol{\theta})(F_{g',a} \mid \boldsymbol{\theta})].$$

By the assumption of homogeneous errors,

$$F_{g,a} \mid \boldsymbol{\theta} \equiv F_{g,a'} \mid \boldsymbol{\theta}, \quad F_{g',a} \mid \boldsymbol{\theta} \equiv F_{g',a'} \mid \boldsymbol{\theta}.$$

Hence the expectation  $\mathcal{E}_a$  is taken over a constant, and therefore

$$F_{gg',*} \mid \boldsymbol{\theta} = (F_{g,a} \mid \boldsymbol{\theta})(F_{g',a} \mid \boldsymbol{\theta}),$$

or

$$F_{gg',*} \mid \boldsymbol{\theta} = (F_{g,*} \mid \boldsymbol{\theta})(F_{g',*} \mid \boldsymbol{\theta}).$$

Thus the assumption of conditional independence is implied by the joint assumptions of experimental independence and homogeneity of errors.

The converse is also true: Given conditional independence,

$$\mathcal{E}_{a \mid \boldsymbol{\theta}} F_{gg',a} \equiv F_{gg',*} \mid \boldsymbol{\theta} = (F_{g,*} \mid \boldsymbol{\theta})(F_{g',*} \mid \boldsymbol{\theta}) = (\mathcal{E}_{a \mid \boldsymbol{\theta}} F_{g,a})(F_{g',a}).$$

This statement must hold for all  $(X_g, X_{g'})$ , and for any subset of  $\mathcal{P}$ . In particular, it must hold for all pairs of  $a \in \mathcal{P}$ . If this expectation equation is to hold for arbitrary subsets (including pairs), then the values taken as a function of  $(X_g, X_{g'})$  must be constant for all  $a$ . Hence

$$F_{gg',a} | \boldsymbol{\theta} = F_{gg',a'} | \boldsymbol{\theta},$$

and hence

$$F_{g,a} | \boldsymbol{\theta} = F_{g,a'} | \boldsymbol{\theta}.$$

Thus the assumption of conditional independence implies the assumption of homogeneous errors. Furthermore

$$F_{gg',a} | \boldsymbol{\theta} = (F_{g,a} | \boldsymbol{\theta})(F_{g',a} | \boldsymbol{\theta}), \quad \text{or} \quad F_{gg',a} = F_{g,a}F_{g',a}, \quad \text{in } \mathcal{P},$$

and hence the assumption of conditional independence implies the assumption of experimental independence. Thus we may state

**Theorem 24.5.3.** The assumption of strong conditional independence is equivalent to the assumptions of experimental independence and homogeneity of errors, taken jointly.

It is important to note that in application, the tenuous part of the assumption of this theorem is satisfied vacuously. Latent trait models, which it may sometimes be useful to characterize by (24.5.1), are typically applied to a single replication of each of the measurements  $g$ . Thus, while it is necessary to assume experimental independence among these measurements, the less desirable assumption of experimental independence between replications of the same measurement is nowhere required. Also, for most practical purposes, it is sufficient to assume that (24.5.1) holds in  $\mathcal{P}$  but not necessarily in every subpopulation of  $\mathcal{P}$ . We made similar comments about the concept of parallelism in Chapter 2. Finally we may note that experimental independence is actually a kind of conditional independence in which the conditioning variable is the person rather than the latent trait.

Given this definition of local independence, we may suggest one possible explication of the concept of a latent trait model and the special case of a single trait model.

**Definition 24.5.4.** Measurements  $X_1, X_2, \dots, X_n$  satisfy a latent trait model if (24.5.1) holds for some *vector-valued trait*  $\boldsymbol{\theta}$ .

**Definition 24.5.5.** Measurements  $X_1, X_2, \dots, X_n$  satisfy a single-trait model if (24.5.1) holds for some *real-valued trait*  $\theta$ .

In Section 24.6, we shall note that according to this definition, the *factor analytic* model as specified in Section 24.3 is a *latent trait* model only if the additional assumption of local independence is introduced. It is clear that these definitions are primarily of theoretical interest.

## 24.6 A Relationship between True Scores and Latent Traits

In studying tests that measure a single underlying trait, in particular, but also in studying tests in general, it is of interest to determine the conditions under which a simple relationship exists between the underlying trait and the true score. Although this question leads directly to a number of very difficult problems, some of which are as yet unsolved, it will be useful to present one definition that explicates this general notion, together with a simple theorem, and to discuss both the practical and theoretical importance of the definition and theorem.

Assume a single underlying real-valued (ability) parameter  $\theta$ , and let  $F_\theta(x)$ , for fixed  $\theta$ , be the conditional distribution function of the observed score. The distribution function  $F_\theta(x)$  characterizes the relative frequency of observed scores over repeated sampling, with replacement, in the subpopulation  $\mathcal{P}_\theta$  of persons having the particular trait value  $\theta$ , or in any proper subpopulation of  $\mathcal{P}_\theta$ . For concreteness, we may think of  $\theta$  as representing verbal ability. Now if an examinee receives an observed score of 90 on a verbal ability test and a second examinee receives a score of 110 on this same test, then we would like to consider this as evidence that the first examinee's verbal ability is less than the second examinee's. Of course, we recognize that the obtained scores differ from the true scores because of (possibly substantial) errors of measurement; nevertheless we would hope that our measurement procedure is such that the results provide some evidence in the indicated direction. Such an interpretation is justified in the context of *any* current theory of statistical inference, provided that the family of cumulative distribution functions  $F_\theta(x)$  is stochastically increasing.

**Definition 24.6.1.** A real-parameter family of distributions  $F_\theta(x)$  is said to be *stochastically increasing* if the distributions are distinct and if  $\theta_1 < \theta_2$  implies  $F_{\theta_1}(x) \geq F_{\theta_2}(x)$  for all  $x$ . The same term is applied to random variables possessing these distributions.

This definition is a technical statement of the condition that for arbitrary values of  $\theta_1$  and  $\theta_2$ , the corresponding distribution functions do not cross, and that the function for  $\theta_1$  is in fact never below the function for  $\theta_2$ .

Now  $F_\theta(x)$  is just

$$\text{Prob}(X \leq x | \theta) = 1 - \text{Prob}(X > x | \theta).$$

Thus, if  $F$  is stochastically increasing,  $\text{Prob}(X > x | \theta)$  is a strictly increasing function of  $\theta$  and  $\text{Prob}(X \leq x | \theta)$  is a strictly decreasing function of  $\theta$ . Hence a high value of  $X$  suggests a high value of  $\theta$ , and a low value of  $X$  suggests a low value of  $\theta$ .

An important feature of a stochastically increasing family is given in

**Lemma 24.6.2.** If  $F_\theta(x)$  is stochastically increasing in  $\theta$ , then  $\mathcal{E}(X | \theta)$  is a strictly increasing function of  $\theta$ .

The proof of this lemma follows immediately from Definition 24.6.1. For distributions not having the property that the regression curve  $\mathcal{E}(X | \theta)$  is a strictly increasing function of  $\theta$ , large values of  $X$  do not necessarily suggest large values of  $\theta$ ; indeed, they may strongly suggest intermediate or low values of  $\theta$ . The inverted U-shaped regression curve often found when task performance is related to frustration is an instance in which a low value of  $X$  would strongly suggest either a very high or a very low value of  $\theta$ , but not an intermediate value. Such a performance task could not be taken as a measure of frustration (see the definition of measure given in Chapter 1), although the observed frustration score may well have a monotonic relationship to frustration.

Now that we have Definition 24.6.1 and Lemma 24.6.2, we may give at least a partial answer to an important question about the relationship between true scores and latent traits. The question is: Given an *arbitrary* strictly increasing transformation on the scale of the observed random variable, under what conditions may the scale of the latent variable be transformed by a strictly increasing transformation so that the new latent variable is the true score (expected value) of the transformed observed variable? As a preliminary, we state an obvious generalization of Lemma 24.6.2:

**Lemma 24.6.3.** If  $F_\theta(x)$  is stochastically increasing in  $\theta$ , and  $\psi = \psi(x)$  is a strictly increasing function of  $X$  such that  $\mathcal{E}[\psi(x) | \theta] < \infty$  for all  $\theta$ , then  $\mathcal{E}[\psi(x) | \theta]$  is a strictly increasing function of  $\theta$ .

This leads to

**Theorem 24.6.4.** If  $F_\theta(x)$  is stochastically increasing in  $\theta$ , then, for any transformation  $\psi = \psi(x)$  for which  $\mathcal{E}[\psi(x) | \theta]$  exists for all  $\theta$ , there exists a strictly increasing transformation  $\varphi = \varphi(\theta)$  such that

$$\mathcal{E}[\psi(x) | \varphi] = \varphi.$$

*Proof.* Let  $\varphi = \varphi(\theta) = \mathcal{E}[\psi(x) | \theta]$ . By Lemma 24.6.3, this is a strictly increasing function of  $\theta$ .  $\square$

The random variable  $\varphi = \varphi(\theta)$  is then the true-score random variable corresponding to the observed-score random variable  $\psi = \psi(x)$ . Now each variable of a set of arbitrarily scaled observed variables  $\mathbf{x}$  may satisfy a single-trait model with the same trait  $\theta$ ; in general, however, it will not be true that a single transformation  $\varphi = \varphi(\theta)$  exists such that  $\varphi$  is the true score for each  $X$ . However, we do have the multivariate generalization of Theorem 24.6.4:

**Theorem 24.6.5.** If  $\mathbf{x}$  is a vector random variable such that each element of  $\mathbf{x}$  satisfies a single-trait model with the common-trait  $\theta$  that stochastically orders each of the variables  $x_i$ , then there exists a set of strictly increasing transformations  $\psi_i = \psi_i(x_i)$ , and a single strictly increasing transformation  $\varphi = \varphi(\theta)$ , such that  $\varphi$  is the true-score random variable corresponding to each element of the vector  $\{\psi_i(x_i)\}$ .

## 24.7 A General Latent Trait Model

Anderson (1959) has given a very general formulation of latent trait theory. From this general model, the logistic and normal ogive response models, a factor analytic model, and many models not considered in this book can be obtained as special cases. As we present this model, we shall follow Anderson's development closely, omitting, however, some work that is not of immediate interest to us.

Let  $\mathbf{x}$ , having  $n$  components, denote the *manifest* variables, and let  $\boldsymbol{\theta}$ , having  $k$  components, denote the latent variables. The number of components of  $\boldsymbol{\theta}$  is less than of  $\mathbf{x}$ , and, hopefully, much less. Also suppose that in some population of people,  $\mathbf{x}$  has a distribution  $H(\mathbf{x} | \boldsymbol{\theta})$ , conditional on  $\boldsymbol{\theta}$ , and that  $\boldsymbol{\theta}$  has a distribution  $G(\boldsymbol{\theta})$ . The unconditional distribution of  $\mathbf{x}$  is then

$$F(\mathbf{x}) \equiv \int H(\mathbf{x} | \boldsymbol{\theta}) dG(\boldsymbol{\theta}). \quad (24.7.1)$$

Typically  $H$  and  $G$  are taken to have densities  $h$  and  $g$ . Then the density of  $F$  is

$$f(x) = \int h(\mathbf{x} | \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (24.7.2)$$

Here, as elsewhere, the term *density function* is taken to refer to the usual density function of a continuous random variable or to the probability mass function of a discrete random variable. The indicated integrals are to be interpreted as Stieltjes integrals; this simply means that when the integration is taken with respect to a continuous variable, the integral is the usual sort of (Riemann) integral, and when integration is taken with respect to a discrete variable, the integration is simply a summation. This same formulation was given in Section 22.3.

The basic assumption of latent trait theory, as we have said, is the assumption of local (conditional) independence (see Sections 16.3, 17.1, 24.5). Under this assumption, it is possible to specify  $\boldsymbol{\theta}$  so that

$$H(\mathbf{x} | \boldsymbol{\theta}) = \prod_{i=1}^n H_i(x_i | \boldsymbol{\theta}). \quad (24.7.3)$$

Some results on moments are available without further assumptions. First we note that since  $\boldsymbol{\theta}$  is unobservable, then

$$\mathcal{E}(\mathbf{x} | \boldsymbol{\theta}) = \boldsymbol{\mu}_{\mathbf{x}|\boldsymbol{\theta}}, \quad (24.7.4)$$

say, is also unobservable; however,  $\mathcal{E}\mathbf{x} = \mathcal{E}[\mathcal{E}(\mathbf{x} | \boldsymbol{\theta})]$  is observable. The conditional matrix of the second-order moments is

$$\mathcal{E}[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|\boldsymbol{\theta}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|\boldsymbol{\theta}})' | \boldsymbol{\theta}] = \mathbf{D}_{\boldsymbol{\theta}}, \quad (24.7.5)$$

say. By (24.7.3),  $D(\boldsymbol{\theta})$  is diagonal. From (24.7.5), we have

$$\mathcal{E}(\mathbf{x}\mathbf{x}' | \boldsymbol{\theta}) = \mathbf{D}_{\boldsymbol{\theta}} + \boldsymbol{\mu}_{\mathbf{x}|\boldsymbol{\theta}} \boldsymbol{\mu}_{\mathbf{x}|\boldsymbol{\theta}}'. \quad (24.7.6)$$

Let

$$\bar{\mathbf{D}} \equiv \|\bar{d}_{ii}\| = \mathcal{E}\mathbf{D}_\theta, \quad (24.7.7)$$

which is diagonal. Then

$$\mathcal{E}\mathbf{x}\mathbf{x}' = \bar{\mathbf{D}} + \mathcal{E}\boldsymbol{\mu}_{\mathbf{x}|\theta}\boldsymbol{\mu}_{\mathbf{x}|\theta}'. \quad (24.7.8)$$

In particular, in an obvious notation,

$$\mathcal{E}x_i^2 = \bar{d}_{ii} + \mathcal{E}\mu_{i|\theta}^2, \quad (24.7.9)$$

$$\mathcal{E}x_i x_j = \mathcal{E}\mu_{i|\theta} \mu_{j|\theta}. \quad (24.7.10)$$

Thus, in principle, one can determine the expected value of  $\boldsymbol{\mu}_{\mathbf{x}|\theta}$  and the covariances of the components of  $\boldsymbol{\mu}_{\mathbf{x}|\theta}$ , but not the variances of these components.

Let  $X_i, X_j, X_k, X_l, \dots$  be distinct random variables. Then a development analogous to that of the previous paragraph establishes that

$$\mathcal{E}x_i x_j x_k x_l \dots = \mathcal{E}\mu_{i|\theta} \mu_{j|\theta} \mu_{k|\theta} \mu_{l|\theta} \dots, \quad (24.7.11)$$

which is observable, and that all variances and mixed higher-order moments of the  $x_i$ , in which any variable has an exponent other than zero or unity, are unobservable. We obtained analogous results in Chapter 10; the reader might see Eqs. (10.5.5) and (10.6.2) in particular.

We may define various special models by introducing further assumptions. We have previously referred to the function  $\mathcal{E}(x | \boldsymbol{\theta})$  as the item characteristic function, or as the item characteristic curve when  $\boldsymbol{\theta}$  is one-dimensional (see Sections 16.1 and 17.1). Perhaps the simplest assumption that may be considered is the assumption of linearity of regression:

$$\mathcal{E}(x | \boldsymbol{\theta}) = \boldsymbol{\mu} + \Lambda\boldsymbol{\theta}. \quad (24.7.12)$$

This has been previously encountered as Eq. (24.2.5). This assumption defines the factor analytic model of Section 2.4, which, with the additional assumption of local independence, we may now consider a latent trait model. This assumption leads to the computation of moments given in Eqs. (24.2.4) and (24.2.9). The assumption of local independence adopted here is a strengthening of the earlier assumption (Eq. 24.2.3) that the components of the residual variables are uncorrelated with each other. In the factor analytic model, we assumed that observed variables  $\mathbf{x}$  and the trait variables  $\boldsymbol{\theta}$  took on continuous values. This assumption is satisfactory even when the observed test score is discrete but takes on a large number of values. For example, a test score obtained by adding zero-one scores from 50 true-false items can typically be treated as continuous.

Here we may draw a useful analogy with the developments in Sections 2.10 and 2.11. We found that because the classical test theory model deals only with first- and second-order moments, the assumption of linear experimental independence can be substituted for the stronger assumption of experimental

independence. Similarly, because the factor analytic model deals only with second-order moments, an assumption of conditional linear independence can be substituted for the usual conditional independence assumption of the latent trait theory. This assumption may be stated by the equation

$$\mathcal{E}(X_g | X_h, X_i, \dots; \boldsymbol{\theta}) = \mathcal{E}(X_g | \boldsymbol{\theta}) \quad \text{for all } g, h, i, \dots . \quad (24.7.13)$$

We have already introduced this assumption in Section 10.2.1, in connection with the estimation of higher-order moments of the true- and error-score distributions. Now, similarly, on considering the test theoretic concept of parallelism in the language of latent trait theory, it is clear that

$$\mathcal{E}(X_g | \boldsymbol{\theta}) \equiv \mathcal{E}(X_{g'} | \boldsymbol{\theta}) \quad (24.7.14)$$

and

$$\sigma^2(X_g | \boldsymbol{\theta}) \equiv \sigma^2(X_{g'} | \boldsymbol{\theta}) \quad (24.7.15)$$

for parallel measurements  $g$  and  $g'$ . The function  $\sigma^2(X_g | \boldsymbol{\theta})$  is sometimes called the *scedastic function of  $X$  given  $\boldsymbol{\theta}$* .

Specific latent trait models may then be determined by specifying particular forms for the item characteristic function  $\pi_i(\theta)$ . [In previous chapters, this function has been denoted by  $P_i(\theta)$ .] The normal ogive model was considered briefly in Chapter 16, and the logistic model was considered in detail in Chapters 17 through 20. The linear factor analytic models were considered earlier in this chapter. Gibson (1960) and McDonald (1962, 1967) have worked with nonlinear factor analytic models. McDonald's use of orthogonal polynomials is particularly attractive. A strong word of caution must be offered with regard to these methods, however: Any kind of factor analysis of real data involves the kind of curve-fitting problems usually encountered in regression analysis and discussed in Chapter 13. When nonlinear factor analytic methods are used, this problem becomes particularly acute. At present, we have inadequate evidence that factors extracted from a nonlinear factor analysis will "cross-validate". Such investigations are overdue.

## 24.8 Latent Trait Models for Binary Random Variables

It is often desirable to specify latent trait models for individual item measurements. In this case, a model specifying continuous values for  $x$  is unreasonable. Two models that do apply here are the normal ogive and logistic item characteristic curve models defined and discussed in Chapters 16 and 17 through 20. Here, again, one assumes that the latent variable takes continuous values.

To conclude our very brief survey of latent trait models, we introduce a general formulation that applies to zero-one items and that assumes that the trait values are continuous and that the latent trait is unidimensional. Except for the logistic and normal ogive models, which are important special cases,

these models are seldom used in psychological testing; however, they have found substantial application in sociology. The general model is the one developed and presented by Lazarsfeld (1950, 1954, 1958) as part of his latent structure analysis for binary (that is, zero-one) items.

If the  $x_i$  are binary, then we may write

$$\begin{aligned}\text{Prob } \{x_i = 1\} &= \mathcal{E}x_k \equiv \pi_i, \\ \text{Prob } \{x_i = 1, x_j = 1\} &= \mathcal{E}x_i x_j \equiv \pi_{ij}, \\ \text{Prob } \{x_1 = 1, x_2 = 1, \dots, x_n = 1\} &= \mathcal{E}x_1 x_2 \cdots x_p \equiv \pi_{12\dots p},\end{aligned}\tag{24.8.1}$$

say. The conditional probabilities for fixed value of the latent trait may be written

$$\begin{aligned}\text{Prob } \{x_i = 1 | \theta\} &= \pi_i(\theta) \\ \text{Prob } \{x_i = 0 | \theta\} &= 1 - \pi_i(\theta).\end{aligned}\tag{24.8.2}$$

The assumption of conditional independence is

$$\pi(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n \pi_i(x_i | \theta),\tag{24.8.3}$$

from which we obtain

$$\mathcal{E}(x_i | \theta) \equiv \mu_i(\theta) = \text{Prob } (x_i = 1 | \theta) = \pi_i(\theta),\tag{24.8.4}$$

$$\mathcal{E}\{[x_i - \mu_i(\theta)]^2 | \theta\} = \pi_i(\theta)[1 - \pi_i(\theta)],\tag{24.8.5}$$

$$\mathcal{E}\{[x_i - \mu_i(\theta)]^3 | \theta\} = \pi_i(\theta)[1 - \pi_i(\theta)][1 - 2\pi_i(\theta)],\tag{24.8.6}$$

$$\mathcal{E}\{[x_i - \mu_i(\theta)]^4 | \theta\} = \pi_i(\theta)[1 - \pi_i(\theta)][1 - 3\pi_i(\theta) + 3\pi_i^2(\theta)].\tag{24.8.7}$$

But

$$x_i = x_i^2 = x_i^3 = x_i^4,$$

and

$$x_i x_j = x_i^2 x_j = x_i x_j^2,$$

and so forth. Hence

$$\mathcal{E}x_i = \mathcal{E}x_i^2 = \mathcal{E}x_i^3 = \mathcal{E}x_i^4\tag{24.8.8}$$

and

$$\mathcal{E}x_i x_j = \mathcal{E}x_i^2 x_j = \mathcal{E}x_i x_j^2.\tag{24.8.9}$$

Thus the general results obtained in the previous section have a somewhat simpler form for binary observed scores.

Two other latent trait models of particular interest are Lazarsfeld's *latent class model* and *latent distance model*. For the latent class model, one assumes that  $\theta$  takes a finite set of  $m$  values, called *classes*. One also assumes that each individual is classifiable into one of  $m$  classes and that the residual distributions

within each class are identical. This kind of assumption seems more in tune with sociological theory than with psychological theory, and the model has indeed been used primarily in the analysis of sociological data.

Finally, we should briefly describe the latent distance model. Suppose that

$$\pi_i(\theta) = \begin{cases} r_i, & \theta \leq c_i, \\ s_i, & \theta > c_i, \end{cases}$$

with  $s_i > r_i$ , and the items are ordered so that  $c_1 < c_2 < \dots < c_m$ . Then, for a sufficiently low value of  $\theta$ , the examinee is likely to respond incorrectly to all items to the extent that all  $r_i$  are near zero: That is, for  $c_1 < \theta \leq c_2$ , the examinee is likely to respond correctly only to the first item; for  $c_2 < \theta \leq c_3$ , he is likely to respond correctly only to the first two items; and so on. For  $\theta$  sufficiently large, he is likely to respond correctly to all items to the extent that all  $s_i$  are near unity. The successive cutoff points  $c_1, c_2, \dots, c_m$  thus define distinct latent classes and also define a kind of ordering of these classes. If  $r_i = 0$  and  $s_i = 1$  for all  $i$ , then this probabilistic model becomes a Guttman "perfect scale".

## References and Selected Readings

- ANDERSON, T. W., On estimation of parameters in latent structure analysis. *Psychometrika*, 1954, **19**, 1-10.
- ANDERSON, T. W., Some recent results in latent structure analysis. In *Proceedings of the 1954 invitational conference on testing problems*. Princeton, N.J.: Educational Testing Service, 1955, pp. 49-53.
- ANDERSON, T. W., Some scaling models and estimation procedures in the latent class model. In O. Grenander (Ed.), *Probability and statistics*, The Harold Cramér Volume. New York: Wiley, 1959, pp. 9-38.
- ANDERSON, T. W., and H. RUBIN, Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Vol. V. Berkeley: University of California Press, 1956, pp. 111-150.
- BROWNE, M. W., Fitting the factor analysis model. *Research Bulletin 67-2*. Princeton, N.J.: Educational Testing Service, 1967.
- GIBSON, W. A., Nonlinear factors in two dimensions. *Psychometrika*, 1960, **25**, 381-392.
- GREEN, B. F., A general solution for the latent class model of latent structure analysis. *Psychometrika*, 1951, **16**, 151-166.
- GREEN, B. F., Latent structure analysis and its relation to factor analysis. *Journal of the American Statistical Association*, 1952, **47**, 71-76.
- GUTTMAN, L., Simple proofs of relations between the communality problem and multiple correlation. *Psychometrika*, 1957, **22**, 147-157.

- HARMAN, H. H., *Modern factor analysis*. Chicago: University of Chicago Press, 1960.
- HARMAN, H. H., and W. H. JONES, Factor analysis by minimizing residuals (Minres). *Psychometrika*, 1966, **31**, 351–368.
- JÖRESKOG, K. G., On the statistical treatment of residuals in factor analysis. *Psychometrika*, 1962, **27**, 335–354.
- JÖRESKOG, K. G., *Statistical estimation in factor analysis*. Uppsala: Almqvist and Wiksell, 1963.
- JÖRESKOG, K. G., Some contributions to maximum likelihood factor analysis. *Research Bulletin 66–41*. Princeton, N.J.: Educational Testing Service, 1966.
- KAISER, H. F., and J. CAFFREY, Alpha factor analysis. *Psychometrika*, 1965, **30**, 1–14.
- LAWLEY, D. N., The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, Series A, 1940, **60**, 64–82.
- LAZARSFELD, P. F., The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer *et al.* (Eds.), *Measurement and prediction*, Chapter 10. Princeton, N.J.: Princeton University Press, 1950.
- LAZARSFELD, P. F., A conceptual introduction to latent structure analysis. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*, Chapter 7. New York: The Free Press, 1954.
- LAZARSFELD, P. F., Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science*, Vol. III. New York: McGraw-Hill, 1958.
- LAZARSFELD, P. F., Latent structure analysis and test theory. In H. Gulliksen and S. Messick (Eds.), *Psychological scaling: theory and applications*. New York: Wiley, 1960, pp. 83–96.
- LEHMANN, E. L., *Testing statistical hypotheses*. New York: Wiley, 1959.
- LORD, F. M., A theory of test scores. *Psychometric Monograph*, No. 7, 1952.
- LORD, F. M., The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953, **13**, 517–549.
- MCDONALD, R. P., A general approach to nonlinear factor analysis. *Psychometrika*, 1962, **27**, 397–415.
- MCDONALD, R. P., Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika*, 1967, **32**, 77–112.
- MEREDITH, W., Some results based on a general stochastic model for mental tests. *Psychometrika*, 1965, **30**, 419–440.
- NOVICK, M. R., The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, **3**, 1–18.
- RAO, C. R., Estimation and tests of significance in factor analysis. *Psychometrika*, 1955, **20**, 93–111.
- RAO, C. R., Characterization of the distribution of random variables in linear structural relations. *Sankhyā*, Series A, 1966, **28**, 251–260.
- ROZEBOOM, W. W., Linear correlations between sets of variables. *Psychometrika*, 1965, **30**, 57–71.

- SPEARMAN, C., General intelligence, objectively determined and measured. *American Journal of Psychology*, 1904, **15**, 201-293.
- SPEARMAN, C., *The abilities of man*. London: Macmillan, 1927.
- THOMSON, G. H., Hotelling's method modified to give Spearman's *g*. *Journal of Educational Psychology*, 1934, **25**, 366-374.
- THURSTONE, L. L., *Multiple-factor analysis*. Chicago: University of Chicago Press, 1947.



## AUTHOR INDEX

- Abelson, R., 22, 25  
Adams, E., 23, 25  
Adams, J. K., 320, 324  
Adams, P. A., 320, 324  
Aitken, A. C., 233  
Albert, A., 314, 321, 323, 326  
Anderson, C. C., 214, 220  
Anderson, T. W., 214, 220, 262, 280,  
    362, 392, 535, 538, 543, 547  
Andrews, F. C., 149  
Aoyama, H., 204, 220  
Avner, R. A., 210, 221  
Azuma, H., 136, 150, 370, 392
- Baird, D. C., 13, 25  
Baker, F. B., 214, 220  
Barton, D. E., 232, 233, 245, 246, 260  
Beaton, A. E., 262, 280  
Bechtoldt, P., 278, 281  
Behnken, D. W., 259  
Berg, J. A., 304, 324  
Berger, A., 337, 354  
Berkson, J., 400, 420, 422, 423, 459, 478  
Bhattacharya, P. K., 340, 354  
Birnbaum, A., 398, 419, 423, 435, 438,  
    439, 442, 445, 451, 452, 456, 471, 478,  
    480, 492  
Birnbaum, Z. W., 149  
Bishop, C. H., 335, 355  
Blyth, C. R., 512, 528  
Bock, R. D., 214, 220, 370, 392, 422, 423  
Boldt, R., 131, 149  
Boomer, D. S., 326  
Brillinger, D. R., 202, 220  
Brogden, H. E., 277, 281, 340, 341, 370,  
    392  
Brown, C. W., 277, 281  
Brown, J., 314, 324  
Brown, W., 84, 112  
Browne, M. W., 535, 547  
Burket, G. R., 290, 299
- Buros, O. K., 196, 215, 220  
Burt, C., 196, 214, 220  
Bush, R. B., 423
- Caffrey, J., 535, 548  
Cahen, L. S., 257, 259  
Calandra, A., 310, 324  
Campbell, A., 14, 25  
Campbell, D. T., 278, 281, 352, 355  
Campbell, N. R., 23, 26  
Carnap, R., 15, 26, 28, 52  
Carroll, J. B., 305, 308, 324, 348, 349,  
    353, 355  
Chernoff, H., 277, 281, 310, 312, 313,  
    324  
Church, A., Jr., 214, 220  
Cleary, T. A., 273, 281  
Clemans, W., 340, 341  
Cochran, W. G., 281, 420, 424  
Cohen, A. C., 149  
Cole, J. W. L., 214, 220  
Coleman, J. S., 150, 214, 220  
Collier, R. O., Jr., 214, 220  
Conrad, H. S., 381, 392  
Converse, P. E., 25  
Coombs, C. H., 314, 317, 324  
Cornfield, J., 163, 164, 171, 235, 236, 260  
Cramér, H., 202, 220, 262, 281, 419, 424,  
    457, 472, 478  
Cromack, T. R., 252, 260  
Cronbach, L. J., 2, 5, 32, 52, 90, 93, 102,  
    119, 128, 133, 136, 150, 175, 195, 196,  
    209, 214, 220, 221, 234, 260, 274, 277,  
    278, 281, 314, 324, 370, 392, 393  
Crow, E. L., 512, 528  
Cureton, E. E., 90, 102, 137, 150, 214,  
    221, 331, 355
- Danford, M. B., 214, 221  
Darlington, R. B., 335, 355  
Das Gupta, S., 336, 355

- David, F. N., 232, 233, 245, 246, 260  
 Davis, F. B., 324  
 Dayhoff, E., 163, 171, 248, 260  
 De Finetti, B., 315, 316, 317, 325  
 Dicken, C., 272, 281  
 Dobbin, J. E., 352, 355  
 Douglass, B., 325  
 Draper, N. R., 289, 299  
 Dressel, P. L., 90, 102, 314, 325  
 Dwyer, P. S., 288, 299
- Eddington, A. S., 497, 500, 506  
 Edwards, A. L., 281  
 Efroyimson, M. A., 289, 299  
 Elfving, G., 290, 300, 335, 355  
 Erdelyi, A., 516, 528
- Fagot, R. F., 23, 25  
 Feller, W., 489, 492  
 Ferguson, G. A., 100, 102, 197, 205, 221  
 Ferguson, T., 505, 506  
 Fifer, G., 325  
 Finney, D. J., 420, 424  
 Fisher, R. A., 259, 260, 420, 424  
 Fiske, D. W., 278, 279, 281, 352, 355  
 Fleiss, J. L., 214, 221  
 Fortier, J. J., 290, 300  
 Frederiksen, N., 273, 281  
 Freeman, H., 4, 5, 35, 52, 206, 221, 281  
 French, J. L., 252, 260  
 French, J. W., 5, 124, 128, 140, 150,  
     298, 300
- Gaffey, W. R., 499, 500, 506  
 Gaito, J., 214, 221  
 Garside, M. J., 269, 281  
 Garside, R. F., 138, 150, 214, 221  
 Gaylord, R. H., 349, 357  
 Gee, H. H., 275, 282  
 Geisser, S., 214, 221  
 Ghiselli, E. E., 26, 102, 273, 277, 281  
 Gibson, W. A., 545, 547  
 Gilbert, A. C. F., 273, 281  
 Girshick, M. A., 493, 506  
 Glaser, R., 214, 221  
 Glass, G. V., 308, 325  
 Gleser, G. C., 2, 5, 32, 52, 119, 128,  
     150, 175, 195, 209, 214, 220–222, 234,  
     260, 274, 277, 281  
 Goldfarb, N., 214, 221  
 Goodman, L. A., 337, 355
- Gourlay, N., 349, 355  
 Graybill, F. A., 4, 5, 163, 166, 171  
 Green, B. F., Jr., 123, 128, 333, 355, 547  
 Greer, D., 252, 260  
 Grizzle, J. E., 214, 220  
 Guilford, J. P., 309, 325, 349, 355  
 Guion, R. M., 273, 281  
 Gulliksen, H., 5, 29, 52, 60, 80, 89, 92,  
     102, 126, 128, 131–133, 135, 141, 150,  
     314, 325, 333, 355, 381, 393  
 Guttman, L., 29, 52, 87, 90, 94, 95, 102,  
     132, 135, 150, 196, 292, 300, 331, 355,  
     404, 424, 533, 547
- Haggard, E. A., 197, 214, 220  
 Haley, D. C., 381, 393, 399, 424  
 Hall, W. J., 442, 452  
 Halperin, M., 5, 26  
 Harman, H. H., 534, 535, 547  
 Hartigan, J. A., 442, 452  
 Hartley, H. O., 5, 26  
 Hayes, S. P., 346, 355  
 Hays, W. L., 5  
 Henrysson, S., 355  
 Hill, B. M., 162, 171  
 Hirschman, I. I., 494, 506  
 Hoel, P. G., 4, 5, 26  
 Hoffman, P. J., 140, 150, 214, 221  
 Hohn, F. E., 281  
 Holtzman, W. H., 214, 221  
 Hooke, R., 163, 171, 238, 239, 244,  
     247, 248, 254, 259  
 Horn, J. L., 214, 221  
 Horst, P., 3, 5, 272, 273, 279, 282, 289,  
     290, 295, 300, 305, 312, 325, 335, 355  
 Howard, K. I., 140, 150, 214, 221  
 Hoyt, C., 90, 102, 197  
 Hughes, H. M., 214, 221  
 Hultquist, R. A., 166, 171  
 Hutchinson, D. W., 512, 528
- Ikeda, H., 210, 221  
 Indow, T., 370, 381, 383, 393
- Jackson, D. N., 5  
 Jackson, R. W. B., 100, 102, 197,  
     205, 221  
 Jaspen, N., 353, 356  
 Jones, W. H., 535, 547  
 Jöreskog, K., 218, 221, 534, 535, 548  
 Joshi, S. W., 494, 507

- Kaiser, H. F., 535, 548  
 Keats, J. A., 519, 528  
 Kelley, T. L., 65, 80, 150  
 Kelly, E. L., 2, 5  
 Kempthorne, O., 14, 26, 420, 424  
 Kendall, M. G., 4, 5, 36, 52, 216, 221,  
   226, 232, 233, 245, 246, 260, 262, 282,  
   287, 300, 346, 356, 385, 393, 455, 457,  
   478, 494, 503, 506, 517, 521, 528  
 Kirkham, R. W., 382, 393  
 Knapp, T. R., 259, 260  
 Koch, S., 30  
 Kogan, N., 320, 325  
 Koutsopoulos, C. J., 63, 80, 81, 94,  
   95, 102  
 Kristof, W., 201, 204, 207, 222  
 Kruskal, W. H., 337, 355  
 Kuder, G. F., 90, 91, 102, 245, 256,  
   258, 517, 523  
 Kurth, R., 498, 506
- LaForge, R., 216, 222  
 Lawley, D. N., 146, 150, 216, 218, 222,  
   369, 393, 467, 478, 535, 548  
 Lazarsfeld, P. F., 29, 36, 52, 362, 393,  
   404, 424, 538-539, 546, 548  
 Lehmann, E., 431, 435, 443, 452,  
   548  
 Lev, J., 335, 356  
 Levine, M., 365, 382, 393  
 Lewis, C., 90, 102, 507  
 Lindgren, B. W., 4, 5, 406, 411, 414,  
   424, 428, 431, 435, 438-440, 445,  
   452  
 Lindley, D. V., 28, 52, 505, 506  
 Lindquist, E. F., 197, 214, 222  
 Lindzey, G., 274, 282  
 Linhart, H., 290, 300, 335, 356  
 Linn, R. L., 144, 150  
 Little, J. M., 275, 282  
 Little, K. B., 214, 221  
 Loève, M., 414, 417, 424  
 Loevinger, J., 27, 52, 235, 260, 278,  
   282, 344, 356, 465, 478  
 Lord, F. M., 73, 75, 81, 104, 120, 122,  
   128, 153, 171, 197, 222, 230, 233, 245,  
   249, 260, 300, 305, 308, 314, 325, 331,  
   340, 356, 370, 383, 388, 393, 420, 422,  
   424, 452, 459, 464, 467, 478, 500, 503,  
   506, 508, 514, 519, 524, 526, 527, 548  
 Lubin, A., 273, 282, 288, 301, 335, 357  
 Luce, R. D., 23, 26  
 Lyerly, S. B., 90, 102
- MacEwan, C., 289, 300, 335, 356  
 Madansky, A., 216, 222, 362, 393  
 Magnussen, D., 150  
 Mandeville, G. K., 214, 220  
 Maritz, J. S., 515, 527, 529  
 Massengill, H. E., 314, 321, 323, 326  
 Mattson, D., 305, 325  
 Maxwell, A. E., 216, 218, 222, 438, 439,  
   445, 452  
 McDonald, R. P., 362, 393, 394, 545, 548  
 McHugh, R. B., 214, 222  
 McKie, D., 102, 234, 260  
 McNee, R. C., 214, 221  
 McNemar, Q., 74, 76, 81, 403, 424  
 Medley, D. M., 197, 214, 222  
 Meehl, P. E., 274, 278, 281, 282  
 Melville, S. D., 273, 281  
 Meredith, W., 95, 102, 140, 150, 214,  
   222, 362, 394, 548  
 Merwin, J. C., 314, 324, 370, 393  
 Messick, S., 5, 352, 356  
 Meyers, E. D., Jr., 344, 356  
 Michael, W. B., 5, 140, 150, 325  
 Milholland, J. E., 314, 324  
 Miller, P. M., 326  
 Miller, R. G., Jr., 202, 222  
 Miller, W. E., 25  
 Mises, L. von, 28, 42  
 Mitzel, H. E., 197, 214, 222  
 Mollenkopf, W. G., 233, 273, 282  
 Mood, A. M., 4, 5  
 Moses, L. E., 277, 281  
 Mosier, C. I., 123, 128  
 Mosteller, F., 25, 26, 285, 292, 300,  
   442, 452  
 Myers, C. T., 344, 356
- Nedelsky, L., 326  
 Neyman, J., 438  
 Nicholson, G. E., 290, 300  
 Norman, W. T., 214, 222  
 Novick, M. R., 29, 52, 81, 90, 102, 104,  
   128, 275, 282, 295, 296, 301, 442,  
   452, 493, 507, 548
- Olkin, I., 201, 208, 222, 282, 285, 300  
 Olsen, M. A., 333  
 Osburn, H. G., 260, 273, 282
- Parratt, L. G., 13, 26  
 Parzen, E., 104, 128

- Paterson, J. J., 370, 394  
 Patil, G. P., 494, 507  
 Paulson, E., 149  
 Pearson, E., 438  
 Pearson, K., 337, 346, 353, 356  
 Peel, E. A., 123, 128  
 Peters, C. C., 131, 147, 150, 282  
 Pitcher, B., 291  
 Pitman, E. J. G., 493, 507  
 Plumlee, L. B., 305, 326, 353, 356  
 Pollard, H., 498, 502, 507  
 Potthoff, R. F., 214, 222  
 Pratt, J. W., 201, 208, 222, 282, 285, 300  
 Raiffa, H., 344, 356  
 Rainwater, J. H., 325  
 Rajaratnam, N., 150, 175, 195, 197,  
   209, 210, 214, 221, 222, 234, 260  
 Rao, C. R., 214, 222, 420, 424, 443, 452,  
   535, 548  
 Rasch, G., 402, 420, 424, 480, 481, 483,  
   485, 486, 489–492  
 Reiersøl, O., 14, 26  
 Richardson, M. W., 90, 91, 102, 245,  
   256, 258, 356, 517, 523  
 Rider, P. R., 282  
 Riordan, J., 232, 233  
 Robbins, H., 494, 507, 527, 529  
 Roberts, A. O. H., 305, 326  
 Robinson, R. E., 23, 25  
 Robson, D. S., 260  
 Rosen, A., 282  
 Ross, J., 346, 356  
 Roy, S. N., 214, 222  
 Rozeboom, W. W., 216, 222, 282, 300,  
   356, 535, 548  
 Rubin, H., 535, 547  
 Rulon, P. J., 90, 102  
 Russell, J. T., 275–278, 283, 301  
 Rydberg, S., 147, 150, 290, 301  
 Samejima, F., 370, 381, 383, 393  
 Saunders, D. R., 273, 282  
 Saupe, J. L., 331, 356  
 Savage, L. J., 315, 326, 442, 452, 493,  
   506  
 Sawyer, J., 274, 282  
 Sax, G., 252, 260  
 Scarborough, J. B., 499, 507  
 Schaie, K. W., 214, 222  
 Scheffé, H., 163, 171  
 Schmid, P., 314, 325  
 Schönemann, P., 102, 234, 260  
 Schrader, W. B., 333  
 Sechrest, L., 274, 282  
 Sherriffs, A. C., 326  
 Shoemaker, D. M., 234, 260  
 Shuford, E. H., 314, 321, 323, 326  
 Sitgreaves, R., 290, 300, 335, 344, 355,  
   357, 465, 479  
 Slakter, M. J., 326  
 Slutsky, E., 202  
 Smith, H., 289, 299  
 Solomon, H., 2, 5, 290, 300, 335, 355,  
   465, 479  
 Spearman, C., 84, 112, 150, 372, 536,  
   548  
 Stalnaker, J. M., 326  
 Standish, C., 502, 507  
 Stanley, J. C., 166, 171, 214, 222, 223  
 Stein, C., 290, 301  
 Stevens, S. S., 26  
 Stevens, W. L., 420, 424  
 Stewart, R., 325  
 Stokes, D. E., 25  
 Stouffer, S. A., 331, 357  
 Stricker, L. J., 273, 283  
 Stuart, A., 4, 5, 36, 52, 216, 221, 226,  
   232, 233, 262, 282, 287, 300, 346, 356,  
   385, 393, 455, 457, 478, 494, 503, 506,  
   517, 521, 528  
 Summerfield, A., 288, 301, 335, 357  
 Suppes, P., 23, 26  
 Sutcliffe, J. P., 26, 28, 39, 42, 52  
 Swineford, F., 326, 402  
 Swoyer, V. H., 289, 301  
 Tallis, G. M., 346, 357  
 Tan, W. Y., 162, 172  
 Tate, R. F., 340, 357  
 Taylor, C. W., 295, 301  
 Taylor, H. C., 275–278, 283, 301  
 Teicher, H., 494, 507  
 Thompson, W. A., Jr., 162, 172  
 Thomson, G. H., 73, 81, 123, 128, 535,  
   548  
 Thorndike, R. L., 3, 5, 27, 39, 52, 142,  
   147, 150, 166, 172, 214, 223, 272, 283,  
   301  
 Thurstone, L. L., 275, 283, 309, 326,  
   344, 357, 534, 536, 548  
 Tiao, G. C., 162, 172  
 Tikhonov, A. N., 527, 529  
 Topping, J., 13, 26  
 Torgerson, W. S., 15, 16, 26, 337, 357,  
   394, 404, 420, 424

- Tricomi, F. G., 527, 529  
Trumpler, R. J., 14, 26, 497, 499, 500,  
502, 507  
Tryon, R. C., 49, 52, 197  
Tucker, L. R., 214, 223, 245, 260, 344,  
357, 370, 394, 420, 424, 465, 479, 523,  
529  
Tukey, J. W., 22, 23, 25, 163, 164, 171,  
232, 235, 260, 442, 452  
Turnbull, W. W., 71, 72, 81, 255  
Twomey, S., 527, 529  
United States National Bureau of  
Standards, 346, 357  
Van Naerssen, R. F., 323, 326  
Van Voorhis, W. R., 131, 147, 150, 282  
Vinsonhaler, J. F., 140, 150, 214, 223  
Votaw, D. F., 326  
Wald, A., 457, 479  
Wall, F. J., 214, 222  
Wallace, D., 285, 292, 300, 442, 452  
Wallach, M. A., 320, 325  
Walsh, J. E., 525, 529  
Warrington, W. G., 133, 150, 370, 393  
Weaver, H. F., 14, 26, 497, 499, 500, 502,  
507  
Webster, H. A., 195, 197, 335, 344, 357  
Weitzman, R. A., 346, 356  
Wherry, R. J., 283, 286, 287, 288, 301,  
349, 357  
Widder, D. V., 494, 506  
Wiggins, R. A., 214, 223  
Wiley, D. E., 214, 221, 308, 325  
Wilks, S. S., 236, 259, 260, 262, 283, 326  
Williams, R. H., 304, 308, 326  
Winer, B. J., 5, 163, 172  
Wolf, R., 357  
Womer, J. F. B., 314, 324, 331  
Woodbury, M. A., 64, 104, 108, 120,  
122, 128, 295, 296, 301  
Yates, F., 259, 260  
Ziller, R. C., 326  
Zimmerman, D. W., 304, 308, 326  
Zinnes, J. L., 23, 26



# TOPIC INDEX

- Ability, 3, 13, 19, 392, 397  
estimation of, 405, 453  
high, 436  
low, 436  
maximum likelihood estimation of,  
    420, 458  
parameter, 480, 486  
Absolute measurement, 21, 43  
Absolute scale, 21  
Admissible operation, 17  
Admissible scoring procedures, schematic  
    representation, 440  
    uniqueness for two-point problems, 441  
Admissible transformation, 20  
Alpha, coefficient, *see* Coefficient alpha  
Analysis of variance components, 151,  
    162, 185–186, 190, 259  
Analysis of variance model, 176  
Applicant group, 141  
Aptitude, 3  
Asymptotic efficiency of maximum  
    likelihood estimates of ability, 455  
Asymptotic moments of estimators of  
    ability, 419  
Attenuation, correction for, 69, 137, 213  
Attenuation paradox, 344, 368, 465  
Average standard error of measurement,  
    *see* Standard error of measurement  
Axioms of test theory, *see also* Classical  
    linear model, 24, 28, 29, 36  
  
Backward elimination procedure, 289  
Balanced designs, 259  
Bayes, empirical problems, 493–494, 527  
Bayesian analyses, 2  
Bernoulli trials with variable parameters,  
    485  
Best difficulty levels, 450–451  
    approximate determination of, for a  
        special case, 450  
    exact determination of, 451  
  
Best item weights, 430  
    locally, 442, 443  
Beta distribution of true scores, 520  
Bias of a test, 187  
Binary variables, *see also* Dichotomous  
    items, 97, 335, 338  
Binomial distribution, 250, 406  
Binomial error model, 250, 508, 523–524  
    consistency with classical model, 510  
Biological assay, 420  
Bipolykays, 238, 247, 254  
Biserial correlation coefficient, 337, 353,  
    378, 379, 403  
    comparison with point biserial, 148  
    effect of item difficulty on, 148  
Bivariate distribution of observed  
    scores, 248, 505, 518–519, 527  
Bivariate distribution of true score and  
    observed score, 494, 496, 513  
Bounded observed scores, 509  
Brogden-Clemans estimator, 341  
  
Calibration sample, 285  
Canonical form of a test, *see* Information  
    structure of a test  
Capitalization on chance, 285, 334  
Ceiling effect, 233, 493  
Central limit theorem, application to  
    weighted sums of test scores, 414  
Centroid, 333, 338  
Chance score, 304  
Change, *see* Gain  
Characteristic function, 504  
Classical linear model, 28, 38, 103–104,  
    531  
    alternative assumptions for, 63  
    assumptions of, 55, 105  
    continuous analogue of, 105–108  
    derivation of, 36  
    estimates of parameters of, 151  
    regression theory for, 64

- Classification, 436  
 two-point problems, 437
- Classification by ability level, 405
- Classification rules, 437, 442, 444  
 uniformly best, 441
- Coefficient alpha, *see also*  
 Kuder-Richardson formula-20, 87,  
 90, 93, 204, 211, 331
- Coefficient of equivalence, 137
- Coefficient of equivalence and stability, 137
- Coefficient of precision, 134
- Coefficient of stability, 137
- Common factor, 216, 371, 381
- Communalities, 372, 533, 536
- Component measurements, 82, 327
- Components analysis, 160–166, 185,  
 244, 259
- Composite, optimally weighted, 350
- Composite measurements, *see also*  
 Composite tests, 82, 444  
 moments of locally best, 416  
 most reliable with a specified true  
 score, 119–121, 123
- Composite tests, 82, 97–98, 237  
 correlation with component, 99  
 correlation between two, 98  
 reliability of, 84, 204  
 weighted, *see* Weighted composites
- Compound binomial distribution, 384,  
 524
- Compound distributions, 494
- Conditional independence, *see* Local  
 independence
- Confidence coefficient, associated with  
 interval estimate of ability, 409
- Confidence interval estimates of true  
 score, 159
- Confidence interval estimates of ability,  
 511–512
- Confidence limit estimator of ability,  
 409
- Consistent estimate, 192, 202  
 of error variance, 192  
 of reliability, 202
- Constitutive definitions, 15
- Construct validity, 261, 278, 352
- Convergent validity, 279, 352
- Convolutions, 494
- Correction for attenuation, *see* Attenuation, correction for
- Correction for chance success, 307,  
 352–353
- Correction for random guessing, 307,  
 352–353
- Correction for shrinkage, 287
- Correlated errors, *see* Errors of  
 measurement, correlated
- Correlation, of errors and true scores,  
 182  
 of errors over examinees, 183  
 between errors on two tests, 36, 56  
 between generic errors and true score,  
 191  
 between item and test score, 330, 341  
 between observed and error scores, 57  
 between observed score and true  
 score, 57, 213  
 between parallel forms of infinite  
 length, 111  
 between parallel forms of zero length,  
 111  
 between sum and a fixed test, 99  
 between sums or averages, 98  
 between total tests and subtests, 98  
 between true and error scores, 36, 56  
 between true scores, 70  
 between two items, 334, 335, 378–379  
 between weighted sums or averages, 97
- biserial, *see* Biserial correlation  
 coefficient
- coefficient, 265
- correction for attenuation, *see*  
 Attenuation, correction for  
 disattenuated, 115
- effect of explicit selection on, 141
- effect of incidental selection on,  
 144–148
- effect of multivariate selection on,  
 146–147
- maximize by adjusting test lengths,  
 124, 295–298
- multiple, *see* Multiple correlation  
 partial, 266
- point biserial, *see* Point biserial  
 correlation coefficient
- polychoric, 353
- rank, 215
- relation between multiple and partial,  
 267
- spurious, 98, 330
- tetrachoric, *see* Tetrachoric correlation
- triserial, 353
- Correlation matrix, dimensionality of, 382
- Correlation ratio, 263
- Costs, 445

- Covariance between observed scores, 62  
 Criterion, 261, 278, 332  
 Criterion problem, 2  
 Critical score, 446  
 Cross validation, 285, 291, 334, 350  
 Cumulant generating function, use in estimating latent moments, 494, 503  
 Cumulants, 226, 247  
 high-order, 230  
 Cutting score, 390
- Decision theory, 2, 3  
 Delta (College Board measure of item difficulty), 381  
 Density function, 18  
 Deterministic models, 23  
 Dichotomous items, *see also* Binary variables, 97, 303  
 product moment correlations for, 335  
 Differences, *see also* True differences, 187  
 between groups, 257  
 between observed scores, 159, 178, 187, 194  
 between two parallel measurements, 154  
 Differential predictability, 271, 273  
 Differential prediction, 3, 273  
 Difficulty (level), 97, 164, 177, 194, 237, 329, 368, 377, 390, 409  
 Dilution series, 420  
 Dimensionality of a latent space, 359, 382  
 Discriminant validity, 279, 352  
 Discriminating items, 251, 392  
 Discriminating power, 329, 367, 378, 389  
 of a classification rule, 409  
 differences in among items, 402–403  
 of a test at different ability levels, 385  
 Discriminating test, 389  
 Discrimination, 392  
 Distinct measurements, 36, 45, 48, 107, 327  
 Distortions in mental measurement, 387–392  
 Distractors, 309  
 worst, 314  
 Distribution function, 18  
 Distributions of test scores, 414  
 Disutilities of errors of classification rules, 445
- Effect (in analysis of variance), 163–166  
 Efficiency of estimators, 436  
 statistical, 444  
 Efficiency-robust methods of estimation, 425, 437  
 Empirical content of models, 398  
 Empirical validity of a model, 491  
 Epistemic definitions, 15  
 Equating, 527  
 Error, 24  
 of estimation, 66, 67  
 of measurement, 13, 14, 27, 31, 66, 67, 493  
 of measurement, generic, 176  
 of measurement, specific, 177  
 of measurement, standard, *see* Standard error of measurement  
 of prediction, 66, 68  
 variable, 535  
 Error probabilities, 437, 442  
 Error score, construction of, 31  
 correlation with observed score, 59  
 correlation with true score, 36, 56  
 expected value, 36, 56  
 for a fixed person, 31–32  
 Platonic, *see* Platonic error score  
 for a randomly selected person, 33  
 variance, for a fixed person, 32  
 variance, for a randomly selected person, 32  
 Error-score distributions, unbiased estimates of the moments of, 224  
 Error-score moments, as a linear function of observed moments, 228  
 Error variance, *see also* Standard error of measurement, 59, 110, 194, 198, 251, 536  
 comparisons of estimates of, 192–193  
 consistent estimate of, 192  
 for a composite test, 205  
 conventional formulas, 192  
 dependent on true score, 229, 251  
 estimates of, 151  
 estimate of generic for test  $g$ , 189  
 estimate of group generic, 177  
 estimate of group specific, 193  
 estimation of, for a given examinee, 154, 178, 251  
 generic, 177  
 generic, estimate of from two randomly selected test forms, 178  
 generic, for examinee  $A$ , 177, 193

- generic, for group, 194
- generic, for group, for fixed test  $g$ , 193
- generous estimate of specific, 167, 186, 192
- group, 251
- scedastic, 308
- specific, *see also* Specific error variance, 151–155, 179, 193
- specific, for examinee  $A$ , 154, 193
- specific, for examinee  $A$  averaged over tests, 193
- table of estimates of, 193
- in terms of observed-score moments, 226
- topastic, 308
- unbiased estimate in item sampling, 251
- upper bound on estimate of specific, 166–167
- upper bound for group specific, 195
- use of estimated, 159
- Errors of measurement, 13, 66, 67
  - basic properties of generic, 180
  - biased, 180
  - correlated, 181–184
  - correlation with true score, 182, 189
  - dependent on true score, 229, 233, 251, 509
  - distribution of, over randomly parallel forms, 251
  - frequency distribution of, for binary items, 250
  - homoscedastic, 129
  - independent of true score, 493
  - normally distributed, 495, 502–503
  - not independent of true score, 233, 251, 509
  - not normally distributed, 233, 503
  - skewness of, 229, 233
  - theory of, 13
  - unbiased, 226
- Essentially  $\tau$ -equivalent measurements, 50, 135
- Estimates, accuracy of, 134
  - choice among, 194
  - consistent, *see* Consistent estimate of score means, sampling variances for, 253–254
  - unbiased, 151, 243
  - unbiased in matrix sampling, 251
- Estimation, of bias for test  $g$ , 187
  - factors effecting, 129
  - of higher-order moments, 224
- Evidence, statistical, 442
- Exhaustion procedure, 288
- Expectation, conditional, 262
- Expected score, maximization of, 315
- Experimental designs, 253
- Experimental independence, 44–45
- Explication, 16
- Explicit selection, 141
- Explicit selection variable, 141
- Exponential form, 445
- Factor, single common, 216, 371, 379, 381
- Factor analysis, 218, 382
  - lack of independence of test length, 536
- Factor analytic decomposition, 534, 535
- Factor analytic model, 530, 540
- Factor loadings, 531, 534
- Factor structure, 290
- Factors, 530–531
- Flexible scoring methods, 317
- Floor effect, 233, 493
- Formula score, 305, 352
- Forward selection procedure, 288
- Future samples, prediction in, 289
- Gain, 73, 159
- Gamma distributions, 480
- Gaussian distribution, 18
- General exponential class, *see* Koopman-Darmois form
- General intelligence factor, 536
- Generalizability, 175, 214
- Generalized symmetric means (gsm), 236–245
- Generating functions, 481
- Generic error variance, *see also* Errors of measurement, generic
  - bias of, 187–188
  - correlation with true score, 191
  - relation with specific error variance, 179
- Generic parameters describing a single test form, 187–191
- Generic reliability, *see also* Reliability, generic
  - relation to coefficient  $\alpha$ , 246
- Generic reliability coefficient, consistent estimate of, 208
- Generic true score, 43, 174
- Generous estimate, 201, 203

- Graduating observed distributions, 517  
 Group mean differences, estimation of, 257–259  
*gsm*, *see* Generalized symmetric means  
 Guessing, 303, 353, 383  
     correction for, 306  
     after eliminating one or more distractors, 307  
 Guessing score, expected, 307  
 Guttman scale, 337, 403–404, 547
- Heterogeneity of group, effect on correlation, *see also* Correlation, effect of selection on  
 Heterogeneous groups, 199, 390  
 Higher-order moments, estimation of, 224, 246  
 Homogeneous items, 363, 381–382  
 Homogeneous scales, 351  
 Homogeneous tests, 103  
 Homoscedasticity, 229  
 Hypergeometric distribution, *see* Negative hypergeometric distribution
- ICC, *see* Item characteristic curve  
 Impediments, 487  
     additivity of, 487  
     relative, 488  
 Imperfectly parallel measurements, 173  
 Incidental selection, 142  
 Incidental selection variable, 142  
 Incorrect score, 132  
 Incremental validity, 273–274  
 Index of reliability, 61  
 Index of speededness, 133  
 Indicant, 19–20  
 Industrial gauging, 420  
 Information, amount of in an item, 367  
 Information, of a given scoring formula, 453  
 Information function, of a scoring formula, 418  
 Information functions, some algebra of, 453–454  
 Information measure of an item, 449  
 Information structure of a test, 410, 425  
 Inheritance on the average, 240  
 Integral equation, 526  
 Interaction effect, 165, 168, 177, 179  
 Interaction variance, 163, 167, 168  
 Internal analysis, estimate of reliability, 134  
 Interval measurement, 21  
 Interval scales, 21, 302  
 Invariant item parameters, 328, 353  
 Invariants, 342, 348  
 Item, *see also* Component measurements  
     correlation with criterion, 332, 341  
     correlation with latent trait, 377  
     intercorrelation, 329, 332, 378–379  
     multiple-choice, 383  
     validity, 332, 351  
     variance, 329  
 Item analysis, *see also* Test design, 327, 380  
     multiple regression methods, 350  
 Item characteristic curve (ICC), 366, 397, 436  
 Item characteristic functions, 358, 360  
 Item difficulties, uniformly distributed, 466  
 Item difficulty, *see also* Difficulty level, 97, 237, 328, 379  
     parameter, 480, 486  
     wide range of, 392  
 Item discriminating power, 329, 331, 377, 389  
 Item homogeneity, *see also* Homogeneous tests, 363, 381  
 Item information function, 450, 453  
 Item mean score, 328  
 Item parameters, 328, 376  
     basic requirement for, 328  
     estimation of, 420  
     invariant, 353, 359, 376, 379  
 Item population, 235  
 Item response pattern, 362  
 Item responses, moments of, 415  
 Item sampling, 204, 234  
     model, 523  
     as a technique in research design, 252–253  
 Item scoring, 319  
 Item-scoring formulas, 302, 313  
 Item selection, *see also* Test design, 343, 350–352, 468  
 Items, difficult, 392  
     information functions of, 460  
     pool of, 235, 380  
     undiscriminating, 392  
 Item-test correlation, 330  
 Item-test regression, 363  
 Item true-score regression, 387  
 Item-weighting formulas, 302  
 Item weights, 350

- Jackknife, 202
- Knowledge or random guessing model, 303, 312
- Koopman-Darmois form, 445
- Kuder-Richardson formula-20, *see also* Coefficient alpha, 91, 101, 245, 256, 258, 526
- Kuder-Richardson formula-21, 91, 101, 517, 523
- Latent class model, 546
- Latent distance model, 546–547
- Latent linear model, 404
- Latent space, 359  
complete, 359, 361  
one-dimensional, 381
- Latent structure analysis, 362
- Latent trait models, 381–383, 397, 540  
basic equations of, 496  
for binary random variables, 545–547  
a general formulation, 543–545
- Latent traits, 3, 359, 530, 537  
bivariate distribution with observed score, 386–389  
correlation with true score, 70  
a relation with true score, 386
- Latent variables, 338, 530
- Law of large numbers, 31
- Least distance scoring method, 318
- Least squares estimates, 286
- Least squares polynomials, 231, 522
- Length of a test, 82  
effect of doubling on error of measurement, 83  
effect of doubling on error variance, 86  
effect of doubling on observed-score variance, 86  
effect of doubling on reliability, 84  
effect of doubling on true variance, 86
- function of reliability invariant with, 126  
graphs showing effect on reliability, 113  
graphs showing effect on validity, 116–117  
relative adjustment of, to maximize battery validity, 124–125, 295–298
- for a specified reliability, 119  
for a specified validity, 127
- Lengthened tests, 249, 526  
correlations of, 114  
estimation of observed score statistics, 249–250, 526
- expectations, 112  
reliability of, 112  
variances of, 112
- Leptokurtic observed score distributions, 390
- Levels of measurement, 20
- Level of partial knowledge, 315
- Likelihood equation, 205, 456  
for a general logistic test model, 459
- Likelihood function, 205, 455
- Likelihood ratio statistics, 430
- Linear experimental independence, 45–46, 225
- Linear minimum mean squared error regression function, 264
- Linear model, 165, 176  
construction of, 34–35
- Linear prediction function, 263
- Linear regression, of observed scores, 505  
of true score on observed score, 231, 503, 505, 513
- Linear regression function, 65, 264–265
- Linear structural relations, 217
- Local independence, 362, 398, 436, 530, 538, 540  
characterization of, 540
- Locally best weights, logistic model, 444  
normal ogive model, 444  
three-parameter logistic model, 444
- Location parameter, 493
- Logistic distribution function, 399
- Logistic test model, 399–401, 441–442  
all items having the same discriminating powers, 402  
as an approximation to the normal form, 400
- Manifest variables, 530
- Matched data, 195
- Matched test forms, 187
- Matrix, Gramian, 349
- Matrix, non-Gramian, 349
- Matrix sampling, 236–238
- Maximum likelihood confidence limit estimator, 458
- Maximum likelihood estimates with normally distributed scores, 204
- Maximum likelihood estimators, 455

- Mean square, expected, 163  
 Mean squared error, 64–65  
 Mean squared error of prediction, 264  
 Measure, 14, 19  
 Measurement, 16, 17  
     absolute, *see* Absolute measurement  
     as a basis of model construction, 16  
     fallible, 14  
     interval, *see* Interval measurement  
     level of, *see* Levels of measurement  
     nominal, *see* Nominal measurement  
     ordinal, *see* Ordinal measurement  
     ratio, *see* Ratio measurement  
 Measurement precision, factors affecting, 129  
 Measurement procedures, 302  
 Measurement properties of a particular nominally parallel test, 213  
 Measurement theory, 23  
 Measurements, distinct, *see* Distinct measurements  
     repeated, *see* Repeated measurements  
 Minimal sufficiency, of the weighted sum form for the logistic model, 426  
 Minimal sufficient statistic, 426  
 Misinformation, 303  
 Mixtures of distributions, 494  
 Model, 15  
     binomial, 508  
     compound binomial, 524–526  
     item-sampling, 234  
     for misreadings, 481  
     normal ogive, 365–366  
     Rasch's item analysis, 480  
     for reading speed, 491  
     utility of, 383  
 Models, deterministic, *see* Deterministic models  
     probabilistic, *see* Probabilistic models  
 Moderator variables, 271, 273  
 Moment generating functions, 494  
 Moments  
     of error-score distribution, unbiased estimates of, 225–230  
     estimation of, 219, 225  
     factorial, 521  
     higher-order, 219, 230, 247  
     of item sample, 244  
     of locally best composite scores, 416  
     multivariate, 227, 248  
     of true-score distribution, unbiased estimates of, 219, 229, 245–248, 494  
 Monotone likelihood ratio property, 445  
 Most powerful classification rule, 439  
 Multiple correlation, 265–268  
     maximizing as a function of relative test lengths, 295–298  
     three-variable formula for, 266  
 Multiple correlation ratio, 265  
 Multiple differential prediction, 3  
 Multiple regression methods in test construction, 350  
 Negative hypergeometric distribution, 515  
     bivariate, 520  
 Nominal measurement, 20, 43  
 Nominal scale, 20  
 Nominally parallel, 167, 174  
 Normal distribution, mean of a tail of, 338  
 Normal distribution function, 399  
 Normal ogive item characteristic curves, 358–359, 365–371, 383–384  
     sufficient condition for, 359, 374–376  
 Normal ogive model, conditions leading to, 370  
 Normal ogive parameters, item difficulty, 376–377  
     item discriminating power, 377–378  
     practical use of, 379  
 Norms, 508, 528  
 Number of items required to attain specified error probabilities, 448–449  
 Oblique factors, 534  
 Observed average, 328  
 Observed score, mean, 60  
 Observed-score parameters, expressed in terms of true-score parameters, 55  
 Observed score and true score, relation between, 230–232, 250, 391, 493–495, 500–501, 508, 510–512, 522  
 Observed-score variance, 57, 110, 329  
 Observed-score variances, estimation of by item sampling, 259  
 Omissiveness, 304  
 Omit, 303  
 Order effect, 166  
 Order-preserving transformation, 21  
 Ordinal measurement, 21, 214  
 Ordinal scale, 21  
 Orthogonal factors, 534

- Paradoxical events, 14  
 Parallel forms method, 134, 136  
 Parallel measurements, *see also* Parallel tests, 47–50, 151, 211, 224, 327, 330 relationships based on, 58–60  
 Parallel test forms, 47–48, 327, 330 construction of, 380 equality of intercorrelation of, 59 equality of intercorrelations with any test, 59  
 Parallel tests, *see also* Parallel measurements nominally, 174, 186, 195 randomly, 248, 524  
 Parallelism, necessary but not sufficient condition for, 59  
 Parameter-free distribution, 489  
 Partial correlation, 265–269  
 Partial information, 303  
 Partial knowledge, 315–318  
 Partial regression weights, 266, 267  
 Partial variance, 267  
 Partially speeded tests, 132, 384  
 Perfect scale, 337, 403, 547  
 Personal probability approach to item scoring, assumptions underlying, 319  
 Personal probability space, 316  
 Personnel selection, 3  
 Persons effect, 164  
 Phi coefficient, 335, 346, 353 in sample, 336 use in factor analysis, 349, 382  
 Pigeon-hole model, 236  
 Platonic error score, 42  
 Platonic true score, 38, 39, 42, 530  
 Platykurtic distributions of true score, 389  
 Point-biserial correlation coefficient, 335, 348, 353, 378 maximum possible, 340  
 Poisson distribution, 480  
 Poisson limit with variable Bernoulli parameters, 485–496  
 Poisson process models, 384, 480, 520 statistical properties in sample, 336  
 Population of items, 235  
 Power tests, 131–133  
 Practice effect, 166  
 Prediction, 2, 262 factors affecting, 129  
 Prediction in future samples, 289  
 Prediction theory, 3  
 Predictor variables, formal procedures for selecting, 288 screening of, 269 selection of, 284  
 Pretest, 328, 351, 380  
 Probabilistic models, 23, 24  
 Probabilities of erroneous classification, *see also* Error probabilities, 408  
 Probability integral transformation, 411  
 Propensity distribution, 30  
 Pseudo-parallel measurements, 49  
 Psychological traits, 338, 354  
 Purely rank-order scoring methods, 317  
 Quantal response models, 420  
 Random selection of persons, 32–34  
 Randomized classification rule, 439  
 Randomly parallel, 248, 523–524  
 Rank correlation, 215  
 Rank of a matrix, 372  
 Ranks, 215  
 Ratio estimates, consistency of, 202  
 Ratio measurement, 21  
 Ratio scale, 21, 481  
 Reading ability, 480  
 Reduced rank prediction systems, 290  
 Regression, *see also* Linear regression, 262 of item score on test score, 363 of item on true score, 387 multiple, in test construction, 350 of observed score on true score, 65, 513 of one observed score on another, 505–506 of test score on ability, 385 of true score on observed score, 152–153, 230, 494, 497, 500, 513, 520–521  
 Regression coefficient, 65, 264, 265 of error score on observed score, 66 of one parallel measurement on a second, 66 of true score on observed score, 65  
 Regression function, 263 of an item response on ability, 398 mean squared error, 64 of one parallel measurement on another, 66  
 Regression model for scoring items, 310

- Regression theory, for classical model, 64–66
- Regression weights, from one sample used in second sample, 291–292
- Relation between scores on parallel test forms, estimation of, 248–249, 518, 528
- Relation of observed score to true score, estimation of, 230–231, 248, 496–497, 500–502
- Relations, structural, 217
- Relative error variance, 110
- Relative observed score, 48, 237
- Relative observed score variance, 110
- Relative true score, 48, 237
- Reliabilities, comparisons among tests of differing lengths, 118–119
- Reliability, *see also* Reliability coefficient, 61, 175, 347, 481, 536 alternate form, 213 comparison of different methods of obtaining, 137 of a composite, maximized, 123–124 of a composite test, 204 conditions of measurement affecting, 137 confidence interval estimate of, with normally distributed errors, 207 definition, 61 doubling length, effect on, 84 effect of group heterogeneity on, 129–131, 199 effect of relative test lengths on, 293–294 effect of selection on, 129–131 estimating without using parallel forms, 216–218 formula for weighting components to maximize, 123–124 as a function of test length, 113 generic, 208 as a generic concept, 139–140 generic for a single test, 209 group heterogeneity, effect on, 129–131, 199 index of, 61 internal analysis method, 134 invariant, 215 invariant function of, 126 Kuder-Richardson lower bounds for, 91 length of test necessary for specified, 119 maximizing, 344 maximum likelihood estimation with normally distributed scores, 204–206 negative estimates of, 202 odd-even, affected by time limits, 136 of ordinal measurement, 214–215 parallel tests used for estimating, 136 retest, 211 of simple formula score, 308 Spearman-Brown formula for stepped-up, *see* Spearman-Brown formula specific, 331 speeded tests, 136 test conditions, effect on, 133–137 of a test of infinite length, 111 test-retest, 134 of a test of zero length, 111 time limits as affecting, 135 unbiased estimator, 207–208 of an unweighted composite, 122 Reliability coefficient, *see also* Reliability, Coefficient alpha, Kuder-Richardson estimating the specific, 200–201 factor loading as, 276 frequency distribution of estimator, 206–207 generic, 208–209 invariant, 215 lower bound to specific of test  $g$ , 201 specific, 199 unbiased estimator of, 207–208 use of factor loadings as, 276 use and interpretation of, 211 Repeated measurements, 13 Replications, 33, 46–47 Replications space, 47 Reproducing scoring systems, 321 Residual, 24 Residual variance, 264, 266, 267 Response pattern, 362, 397 Response patterns, 436 Response styles, 304 Restriction of range, correction for, 130 Right-wrong dichotomy, 337 Robustness, 24, 425, 437 Rotation of factors, 534
- Sample estimate, 153
- Sample variance, 153
- Sampling, stratified, 235
- Scedastic error variance, 308

- Score, formula, *see also* Formula score, 436, 471  
 as a statistic, 425
- Score distribution, leptokurtic, 390
- Scoring weights, 145, 442, 472–473
- Screening sample, 285
- Selected group, 141
- Selection formulas, accuracy of, 147–148
- Selection of group, 140–142  
 basic assumptions for correction for,  
 in two-variable case, 142–143  
 correlation between incidental and  
 explicit selection variables, 145–146  
 correlation between two incidental  
 selection variables, 145–146  
 effect of use of fallible measures, 508,  
 528  
 explicit, 141  
 incidental, 141  
 multivariate, basic assumptions for,  
 146  
 practical importance of corrections  
 for, 14, 142  
 reliability affected by, 129–131, 199  
 univariate, basic assumptions for  
 three-variable case, 144–145
- Selection of persons, 390
- Selection of predictor variables, 288–289
- Selection ratio, 275
- Semantic definition, 1, 15
- Separation of parameters, 487–489, 491
- Shrinkage, 285, 292, 334, 350
- Signal-to-noise ratio, 118, 198
- Simple ordering of responses, 312
- Simplex, 316
- Single-factor model, 536
- Single-trait model, 540
- Skewness, 229, 233, 390
- Spearman-Brown formula, 84, 87, 90,  
 112–114, 118, 206, 347, 526  
 accuracy of, 139  
 for double length, 84  
 general case, 112  
 graphs illustrating, 113
- Specific error variance, *see also* Error variance, specific  
 relation with generic error variance,  
 179  
 unbiased estimate of, from parallel  
 measurements, 153–158
- Specific factors, 535
- Specific true score, 43, 151, 152
- Specificity, 536
- Speeded tests, 131–133, 253, 384
- Split halves, 135
- Squared error loss estimator, robustness  
 of, 277
- Standard error of estimation, 67
- Standard error of measurement, *see also*  
 Error variance, 60, 67, 198, 251  
 average, 60  
 for a fixed person, 60  
 at a given ability level, 385
- Standard error of prediction, 68
- State, 28
- Statistic, 425
- Statistical efficiency, 436, 472
- Stepped-up reliability, 135, 206–207
- Stepwise procedures, 289
- Stochastic ordering, 406
- Stochastically increasing family of  
 distributions, 541
- Strict least distance scoring methods, 317
- Structural relations, 216
- Sufficiency, 428, 442
- Sufficiency of a statistic, conditions for,  
 428
- Sufficient statistic, 251, 426, 439, 445  
 factorization theorem, 427  
 test with equivalent items, 429  
 logistic test model, 429, 431–434  
 two-point discrimination problems,  
 429
- Suppressor variables, 271–272
- Symmetric reproducing scoring systems,  
 322
- Syntactic definition, 1, 15, 18
- Tau-equivalent measurements, 47, 224,  
 505  
 basic theorem for, 227
- Test, as a measuring instrument, 405,  
 508
- Test characteristic curve, 386
- Test construction, *see also* Test design,  
 327, 350–351, 392  
 empirical approaches to, 350–351
- Test construction techniques,  
 considerations in the choice of,  
 350–352
- Test design, *see also* Item selection,  
 Item analysis, 436, 468
- Test development technology, 3
- Test difficulty, *see* Difficulty
- Test effect, 164
- Test information function, 454

- Test item, 327  
 Test items, factor analyses of, 349, 381–382  
 Test length, *see also* Length of a test, 103–104  
 Test length parameter, 104–105  
 Test models, *see also* Models, 397, 436  
 Test-retest method of estimating reliability, 134  
 Test scores, *see also* Observed scores, 425  
     conditional distribution of, 384  
     quantiles of, 415  
 Test with a wide range of item difficulty, 392  
 Tests, imperfectly parallel, 217  
     information functions of, 465  
 Tetrachoric correlation, 337, 345, 353, 376, 379, 382  
     and comparison with phi coefficients, 346–349, 382  
 Tetrad difference equation, 372  
 Theoretical construct, 261  
 Theoretical validity of a test, 261  
 Theory, psychological, 15  
 Theory of errors, 13, 14  
 Theory of measurement, *see* Measurement theory  
 Theory of psychology, 358  
 Time limit, 132  
 Topastic error variance, 308  
 Trait, 27–28, 358  
     latent, *see* Latent traits  
 Traits, unitary, 351  
 Transformation, order-preserving, *see* Order-preserving transformation  
 Transformation of scores, 216  
 Transformations of scales of ability, 411  
 Triad, 372  
 True change, 74–76  
 True difference, 73, 159  
 True gain, 73  
 True score, 2, 27–31, 55–57, 493, 530  
     bivariate distribution with observed score, 231–232, 248, 385, 494, 497, 500–501, 508, 510, 521–522  
     confidence intervals for, 511–512  
     construction of, 29  
     construction of generic, 174  
     distribution of, 390, 495, 497–498  
     estimating, 152  
     as an expectation, 30–31  
     expectation of, 56–57  
     generic, *see also* Generic true score, 164, 175, 208, 237  
     as a limiting average of observed scores, 28, 108  
     mean, 60  
     mean squared error regression estimate of, 152  
     moments of bivariate distribution with observed score, 522  
     moments of distribution of, 219, 224–226, 245, 494–495  
     Platonic, *see* Platonic true score  
     regression of observed score on, 57, 513–515  
     relation to latent trait, 387  
     relative, 237  
     specific, *see* Specific true score  
 True-score distribution, skewness of, 229, 390  
     U-shaped, 390  
     variances of, 56  
 True-score variance, equal to covariance between parallel measurements, 59  
 estimation of, 151–152, 184–186, 245  
 relationship between generic and specific, 185  
 specific, 185  
 specific, estimated from an analysis of variance components, 160–162  
 Two-point problem, 445  
 Two by two tables, 337, 346, 353, 376  
 Uncompleted texts, 490–492  
 Uniform systems analysis, 365  
 Unique factor, 373, 531  
 Uniqueness, 536  
 Unmatched data, 208, 214  
 Unmatched test forms, 187, 195  
 Unweighted scores, efficiency of, in the logistic models, 472  
 Validity, 59, 481  
     bounded by index of reliability, 72  
     comparisons among tests of differing lengths, 118–119  
     construct, 261, 278, 352  
     convergent, 279, 352  
     as a correlation coefficient, 277–278  
     discriminant, 279, 352  
     effect of explicit selection on, 140–141, 145, 146  
     effect of incidental selection on, 142, 144–146

- effect of relative test lengths on, 293–295  
effect of selection on, 144–146  
effect of test length on, 114–115  
empirical, 351–352  
function of, invariant with length of test, 126  
graph showing effect of test length, 116–117  
improving by graphic methods, 333  
length necessary to attain a specified, 127  
of models, 398  
and the selection ratio, 275  
theoretical, 261  
Validity coefficient, 61–62, 261  
practical utility of, 277  
Variance, conditional, 262  
negative estimate of, 162  
Variance components, *see also*  
Components analysis, Analysis of variance components, 160, 190, 244, 259  
Variance of error scores, *see also* Error variance, 60  
effect of test length on, 110  
relation to interaction, 162–166  
Variance of observed score, 56  
effect of doubling length, 83  
effect of test length on, 110  
Variance of true scores, 56  
effect of doubling length on, 83  
effect of test length on, 110  
relation to variance due to persons, 156  
Variance ratio, statistical properties of estimated, 201  
Variance ratios, 198, 200  
Weighted composites, 96–97  
covariances of, 96  
expectations of, 96  
variances of, 96  
Within-set correlation of residual variates, 535



A standard linear barcode is positioned horizontally across the page. It consists of vertical black bars of varying widths on a white background.

9 781593 119348