BEST TEST DESIGN

BEST TEST DESIGN

Benjamin D. Wright
Mark H. Stone
University of Chicago

MESA PRESS
Chicago
1979

MESA PRESS
5835 Kimbark Ave.
Chicago, IL
60637

Copyright © 1979 by Benjamin D. Wright and Mark H. Stone
All rights reserved

Printed in the United States of America
Library of Congress Catalog Card Number 79-88489

Cover Design: Andrew W. Wright
Graphics and Typesetting: Betty Stonecipher

ACKNOWLEDGEMENTS

This book owes more to Georg Rasch than words can convey. The many months of painstaking and inspired tutoring with which he introduced me to the opportunities and necessities of measurement during 1960 are the foundation on which Mark Stone and I have built. I am also indebted to Jane Loevinger for her common sense during the 1965 Midwest Psychological Association Symposium on "Sample-free Probability Models in Psychological Measurement," to Darrell Bock for encouraging me to talk about "Sample-free Test Construction and Test-free Ability Measurement" at the March, 1967 meeting of the Psychometric Society and to Benjamin Bloom for insisting on "Sample-free Test Calibration and Person Measurement" at the 1967 ETS Invitational Conference on Testing Problems.

Benjamin D. Wright

David Farr initiated an AERA Training Session on "Sample-free Test Calibration and Person Measurement in Educational Resarch" in 1969. Since then, thanks to the support of Richard Jaeger and William Russell, there have been several AERA "presessions." These forums for debate on the theory and practice of Rasch measurement have fostered many instructive collaborations. For these lessons in how to make Rasch measurement useful we are especially indebted to Louis Bashaw, Richard Woodcock, Robert Rentz and Fred Forster.

The long hard work of mathematical analysis, computer programming and empirical verification which makes Rasch measurement not only practical but even easy to use was done with friends in the MESA program at The University of Chicago. This great adventure began in 1964 with Bruce Choppin and Nargis Panchapakesan and continues with David Andrich, Graham Douglas, Ronald Mead, Robert Draba, Susan Bell and Geoffrey Masters. We dedicate our book to them.

Benjamin D. Wright
Mark H. Stone
The University of Chicago
March 30, 1979

FORWARD

0.1 HOW TO USE THIS BOOK

This is a handbook for learning how to do Rasch measurement. We give some theoretical explanations, but we emphasize practice. Chapters 2, 4, 5 and 7 use a small problem to illustrate the application of Rasch measurement in complete detail. To those of you who learn best by doing we recommend going directly to Chapters 2 and 4 and working through the practical problem developed there. Next study the sections of Chapters 5 and 7 that continue the analysis of this problem and then go back to Chapter 1.

Behind the practical procedures of Rasch measurement are their reasons. The methodological issues that motivate and govern Rasch measurement are developed in Chapters 1, 5 and 6. To those of you who like to begin with theory, we recommend reading these chapters before working the problem in Chapters 2 and 4. Finally, if Rasch measurement is entirely new to you, you might want to begin with Section 0.3 of this Forward which is an introduction to the topic given at the October 28, 1967 ETS Invitational Conference on Testing Problems (Wright, 1968).

Section 0.2 reviews the motivation and history of the ideas that culminated in Rasch's *Probabilistic Models for Some Intelligence and Attainment Tests* (1960). The references cited there and elsewhere in this book are focused on work that (1) bears directly on the discussion, (2) is in English and (3) is either readily available in a university library or can be supplied by us.

0.2 MOTIVATION AND HISTORY

Fifty years ago Thorndike complained that contemporary intelligence tests failed to specify "how far it is proper to add, subtract, multiply, divide, and compute ratios with the measures obtained." (Thorndike, 1926, 1). A good measurement of ability would be one "on which zero will represent just not any of the ability in question, and 1, 2, 3, 4, and so on will represent amounts increasing by a constant difference." (Thorndike, 1926, 4). Thorndike had the courage to complain because he believed he had worked out a solution to the problem for his own intelligence test. So did Thurstone (1925).

Thurstone's method was to transform the proportion in an age group passing any item into a unit normal deviate and to use these values as the basis for scaling. Common scale values for different age groups were obtained by assuming a linear relationship between the different scale values of items shared by two or more test forms using the different group means and standard deviations as the parameters for a transformation onto a common scale. Thurstone redid a piece of Thorndike's work to show that his method was better (Thurstone, 1927). His "absolute scale" (1925, 1927) yields a more or less interval scale. But one which is quite dependent on the ability distribution of the sample used. In addition to item homogeniety, the Thurstone method requires the assumption that ability is normally distributed within age groups and that there exist

relevant fixed population parameters for these distributions. Should the specification of population be inappropriate so will the estimated scale values. Should the sampling of intended populations be inadequate in any way so will the estimated scale values. They cannot be invariant to sampling. Samples differing in their ability distributions will produce scale values different in magnitude and dispersion.

Thurstone used the 1925 version of his method for the rest of his life, but the majority of test calibrators have relied on the simpler techniques of percentile ranks and standard scores. The inadequacies of these methods were clarified by Loevinger's 1947 analysis of the construction and evaluation of tests of ability (Loevinger, 1947).

Loevinger showed that test homogeniety and scale monotonicity were essential criteria for adequate measurement. In addition, "An acceptable method of scaling must result in a derived scale which is independent of the original scale and of the original group tested." (Loevinger, 1947, 46). Summing up the test calibration situation in 1947, Loevinger says, "No system of scaling has been proved adequate by the criteria proposed here, though these criteria correspond to the claims made by Thurstone's system." (Loevinger, 1947, 43). As for reliabilities based on correlations, "Until an adquate system of scaling is found, the correlation between tests of abilities, even between two tests of the same ability, will be accidental to an unknown degree." (Loevinger, 1947, 46).

In 1950 Gulliksen concluded his *Theory of Mental Tests* with the observation that

> Relatively little experimental or theoretical work has been done on the effect of group changes on item parameters. If we assume that a given item requires a certain ability, the proportion of a group answering that item correctly will increase and decrease as the ability level of the group changes. . . . As yet there has been no systematic theoretical treatment of measures of item difficulty directed particularly toward determining the nature of their variation with respect to changes in group ability. Neither has the experimental work on item analysis been directed toward determining the relative invariance of item parameters with systematic changes in the ability level of the group tested (Gulliksen, 1950, 392-393).

At the 1953 ETS Invitational Conference on Testing Problems, Tucker suggested that, "An ideal test may be conceived as one for which the information transmitted by each of the possible scaled scores represents a location on some unitary continuum so that uniform differences between scaled scores correspond to uniform differences between test performances for all score levels " (Tucker, 1953, 27). He also proposed the comparison of groups differing in ability as a strong method for evaluating test homogeneity (Tucker, 1953, 25). But the other participants in the conference belittled his proposals as impractical and idealistic.

In 1960 Angoff wrote in his encyclopedia article on measurement and scaling that

> Most of the test scales now in use derive their systems of units from data taken from actual test administrations, and thus are dependent on the performance of the groups tested. When so constructed, the scale has meaning only so long as the group is well defined and has meaning, and bears a resemblance in some fashion to the groups or individuals who later take the test for the particular purposes of selection, guidance, or group evaluation. However, if it is found

> that the sampling for the development of a test scale has not been adequate, or that the group on which the test has been scaled has outlived its usefulness, possibly because of changes in the defined population or because of changes in educational emphases, then the scale itself comes into question. This is a serious matter. A test which is to have continued usefulness must have a scale which does not change with the times, which will permit acquaintance and familiarity with the system of units, and which will permit an accumulation of data for historical comparisons (Angoff, 1960, 815).

And yet the faulted methods referred to and criticized by Loevinger, Gulliksen and Angoff are still widely used in test construction and measurement. This is in spite of the fact that considerable evidence has accumulated in the past twenty-five years that much better methods are possible and practical.

These better methods have their roots in the 19th century psychophysical models of Weber and Fechner. They are based on simple models for what it seems reasonable to suppose happens when a person responds to a test item. Two statistical distributions have been used to model the probabilistic aspect of this event. The normal distribution appears as a basis for mental measurement in Thurstone's Law of Comparative Judgement in the 1920's. The use of the normal ogive as an item response model seems to have been initiated by Lawley and Finney in the 1940's. Lord made the normal ogive the cornerstone of his approach to item analysis until about 1967, when under Birnbaum's influence, he switched to a logistic response model (Lord, 1968).

The logistic distribution was used by biometricians to study growth and mortality rates in the 1920's and Berkson has championed its practical advantages over the normal distribution ever since. These biometric applications were finally picked up, probably through the work of Bradley and Terry in the 1950's, and formulated into a logistic response model for item analysis by Birnbaum (1968) and Baker (1961). Baker developed computer programs for applying logit and probit item analysis and studied their performance with empirical and simulated data (Baker, 1959, 1963).

In all of these approaches to item analysis, however, at least two parameters are sought for each item. Attempts are made to estimate not only an item difficulty, the response ogive's horizontal intercept at probability one-half, but also an item discrimination, the ogive's slope at this intercept. Unfortunately this seemingly reasonable elaboration of the problem introduces an insurmountable difficulty into applying these ideas in practice. There has been a running debate for at least fifteen years as to whether or not there is any useful way by which some kind of estimates of item parameters like item discrimination and item "guessing" can be obtained.

The inevitable resolution of this debate has been implicit ever since Fisher's invention of sufficient estimation in the 1920's and Neymann and Scott's work on the consistency of conditional estimators in the 1940's. Rasch (1968), Andersen (1973, 1977) and Barndorff-Nielsen (1978) each prove decisively that only item difficulty can actually be estimated consistently and sufficiently from the right/wrong item response data available for item analysis. These proofs make it clear that the dichotomous response data available for item analysis can only support the estimation of item difficulty and that attempts to estimate any other individual item parameters are necessarily doomed.

The mathematics of these proofs need not be mastered to become convinced of their

practical implications. Anyone who actually examines the inner workings of the various computer programs advertised to estimate item discriminations and tries to apply them to actual data, will find that the resulting estimates are highly sample dependent. If attempts are made in these computer programs to iterate to an apparent convergence, this "convergence" can only be "reached" by interfering arbitrarily with the inevitable tendency of at least one of the item discrimination estimates to diverge to infinity. In most programs this insurmountable problem is sidestepped either by not iterating at all or by preventing any particular discrimination estimate from exceeding some entirely arbitrary ceiling such as 2.0.

As far as we can tell, it was the Danish mathematician Georg Rasch who first understood the possibilities for truly objective measurement which reside in the simple logistic response model. Apparently it was also Rasch who first applied the logistic function to the actual analysis of mental test data for the practical purpose of constructing tests. Rasch began his work on psychological measurement in 1945 when he standardized a group intelligence test for the Danish Department of Defense. It was in carrying out that item analysis that he first "became aware of the problem of defining the difficulty of an item independently of the population and the ability of an individual independently of which items he has actually solved." (Rasch, 1960, viii). By 1952 he had laid down the basic foundations for a new psychometrics and worked out two probability models for the analysis of oral reading tests. In 1953 he reanalyzed the intelligence test data and developed the essentials of a logistic probability model for item analysis.

Rasch first published his concern about the problem of sample dependent estimates in his 1953 article on simultaneous factor analysis in several populations (Rasch, 1953). But his work on item analysis was unknown in this country until the spring of 1960 when he visited Chicago for three months, gave a paper at the Berkeley Symposium on Mathematical Statistics (Rasch, 1961), and published *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960).

In her 1965 review of person and population as psychometric concepts Loevinger wrote,

> Rasch (1960) has devised a truly new approach to psychometric problems . . . He makes use of none of the classical psychometrics, but rather applies algebra anew to a probabilistic model. The probability that a person will answer an item correctly is assumed to be the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. Beyond specifying one person as the standard of ability or one item as the standard of difficulty, the ability assigned to an individual is independent of that of other members of the group and of the particular items with which he is tested; similarly for the item difficulty . . . Indeed, these two properties were once suggested as criteria for absolute scaling (Loevinger, 1947); at that time proposed schemes for absolute scaling had not been shown to satisfy the criteria, nor does Guttman scaling do so. Thus, Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of nonarbitrary measures. Rasch is concerned with a different and more rigorous kind of generalization than Cronbach, Rajaratnam, and Gleser. When his model fits, the results are independent of the sample of persons and of the particular items within some broad limits. Within these limits, generality is, one might say, complete (Loevinger, 1965, 151).

0.3 AN INTRODUCTION TO THE MEASUREMENT PROBLEM

My topic is a problem in measurement. It is an old problem in educational testing. Alfred Binet worried about it 60 years ago. Louis Thurstone worried about it 40 years ago. The problem is still unsolved. To some it may seem a small point. But when you consider it carefully, I think you will find that this small point is a matter of life and death to the science of mental measurement. The truth is that the so-called measurements we now make in educational testing are no damn good!

Ever since I was old enough to argue with my pals over who had the best IQ (I say "best" because some thought 100 was perfect and 60 was passing), I have been puzzled by mental measurement. We were mixed up about the scale. IQ units were unlike any of those measures of height, weight, and wealth with which we were learning to build a science of life. Even that noble achievement, 100 percent, was ambiguous. One hundred might signify the welcome news that we were smart. Or it might mean the test was easy. Sometimes we prayed for easier tests to make us smarter.

Later I learned one way a test score could more or less be used. If I were willing to accept as a whole the set of items making up a standardized test, I could get a relative measure of ability. If my performance put me at the eightieth percentile among college men, I would know where I stood. Or would I? The same score would also put me at the eighty-fifth percentile among college women, at the ninetieth percentile among high school seniors, and above the ninety-ninth percentile among high school juniors. My ability depended not only on which items I took but on who I was and the company I kept!

The truth is that a scientific study of changes in ability—of mental development—is far beyond our feeble capacities to make measurements. How can we possibly obtain quantitative answers to questions like: How much does reading comprehension increase in the first three years of school? What proportion of ability is native and what learned? What proportion of mature ability is achieved by each year of childhood?

I hope I am reminding you of some problems which afflict present practice in mental measurement. The scales on which ability is measured are uncomfortably slippery. They have no regular unit. Their meaning and estimated quality depend upon the specific set of items actually standardized and the particular ability distribution of the children who happened to appear in the standardizing sample.

If all of a specified set of items have been tried by a child you wish to measure, then you can obtain his percentile position among whatever groups of children were used to standardize the test. But how do you interpret this measure beyond the confines of that set of items and those groups of children? Change the children and you have a new yardstick. Change the items and you have a new yardstick again. Each collection of items measures an ability of its own. Each measure depends for its meaning on its own family of test takers. How can we make objective mental measurements and build a science of mental development when we work with rubber yardsticks?

The growth of science depends on the development of objective methods for transforming observation into measurement. The physical sciences are a good example. Their basis is the development of methods for measuring which are specific to the measurement intended and independent of variation in the other characteristics of the objects measured

or the measuring instruments used. When we want a physical measurement, we seldom worry about the individual identity of the measuring instrument. We never concern ourselves with what objects other than the one we want to measure might sometime be, or once have been, measured with the same instrument, It is sufficient to know that the instrument is a member in good standing of the class of instruments appropriate for the job.

When a man says he is at the ninetieth percentile in math ability, we need to know in what group and on what test before we can make any sense of his statement. But when he says he is five feet eleven inches tall, do we ask to see his yardstick? We know yardsticks differ in color, temperature, compositions, weight—even size. Yet we assume they share a scale of length in a manner sufficiently independent of these secondary characteristics to give a measurement of five feet eleven inches objective meaning. We expect that another man of the same height will measure about the same five feet eleven even on a different yardstick. I may be at a different ability percentile in every group I compare myself with. But I am the same 175 pounds in all of them.

Let us call measurement that possesses this property "objective". Two conditions are necessary to achieve it. First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for measuring. In practice, these conditions can only be approximated. But their approximation is what makes measurement objective.

Object-free instrument calibration and instrument-free object measurement are the conditions which make it possible to generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to combine or partition instruments to suit new measurement requirements.

The guiding star toward which models for mental measurement should aim is this kind of objectivity. Otherwise how can we ever achieve a quantitative grasp of mental abilities or ever construct a science of mental development? The calibration of test-item difficulty must be independent of the particular persons used for the calibration. The measurement of person ability must be independent of the particular test items used for measuring.

When we compare one item with another in order to calibrate a test, it should not matter whose responses to these items we use for the comparison. Our method for test calibration should give us the same results regardless of whom we try the test on. This is the only way we will ever be able to construct tests which have uniform meaning regardless of whom we choose to measure with them.

When we expose persons to a selection of test items in order to measure their ability, it should not matter which selection of items we use or which items they complete. We should be able to compare persons, to arrive at statistically equivalent measurements of ability, whatever selection of items happens to have been used—even when they have been measured with entirely different tests.

Exhortations about objectivity and sarcasm at the expense of present practices are easy. But can anything be done about the problem? Is there a better way? In the old way of doing things, we calibrate a test item by observing how many persons in a standard

sample succeed on that item. The traditional item "difficulty" is the proportion of correct responses in some standardizing sample. Item quality is judged from the correlation between these item responses and test scores. Person ability is a percentile standing in the same "standard" sample. Obviously this approach leans very heavily on assumptions concerning the appropriateness of the standardizing sample of persons.

A quite different approach is possible, one in which no assumptions need be made about the ability distribution of the persons used. This new approach assumes instead a very simple model for what happens when any person encounters any item. The model says simply that the outcome of the encounter shall be taken to be entirely governed by the difference between the ability of the person and the difficulty of the item. Nothing more. The more able the person, the better their chances for success with any item. The easier the item, the more likely any person is to solve it. It is as simple as that.

But this simple model has surprising consequences. When measurement is governed by this model, it is possible to take into account whatever abilities the persons in the calibration sample happen to demonstrate and to free the estimation of item difficulty from the particulars of these abilities. The scores persons obtain on the test can be used to remove the influence of their abilities from the estimation of item difficulty. The result is a sample-free item calibration.

The same thing can happen when we measure persons. The scores items receive in whatever sample happens to provide their calibrations can be used to remove the influence of item difficulty from the estimation of person ability. The result is a test-free person measurement.[1]

[1] Adapted from *Proceedings of the 1967 Invitational Conference on Testing Problems.* Copyright © 1968 by Educational Testing Service. All rights reserved. Reprinted by permission.

CONTENTS

ACKNOWLEDGEMENTS v

FORWARD vii

1 THE MEASUREMENT MODEL 1
- 1.1 How Tests are Used to Measure 1
- 1.2 How Scores are Used 4
- 1.3 What Happens When a Person Takes an Item 9
- 1.4 The Rasch Model 15
- 1.5 Using the Rasch Model for Calibrating and Measuring 17
- 1.6 A Simple Useful Estimation Procedure 21
- 1.7 How Traditional Test Statistics Appear in Rasch Measurement 24

2 ITEM CALIBRATION BY HAND 28
- 2.1 Introduction 28
- 2.2 The Knox Cube Test 28
- 2.3 The Data for Item Analysis 29
- 2.4 Calibrating Items and Measuring Persons 30
- 2.5 Discussion 44

3 ITEM CALIBRATION BY COMPUTER 46
- 3.1 Introduction 46
- 3.2 BICAL Output for a PROX Analysis of the Knox Cube Test Data 46
- 3.3 Comparing PROX by Hand with PROX by Computer 55
- 3.4 Analyzing KCT with the UCON Procedure 56
- 3.5 Comparing UCON to PROX with the KCT Data 60
- 3.6 A Computing Algorithm for PROX 61
- 3.7 The Unconditional Procedure UCON 62

4 THE ANALYSIS OF FIT 66
- 4.1 Introduction 66
- 4.2 The KCT Response Matrix 66
- 4.3 The Analysis of Fit by Hand 69
- 4.4 Misfitting Person Records 76
- 4.5 Misfitting Item Records 77
- 4.6 Brief Summary of the Analysis of Fit 79
- 4.7 Computer Analysis of Fit 80

5 CONSTRUCTING A VARIABLE 83
- 5.1 Generalizing the Definition of a Variable 83
- 5.2 Defining the KCT Variable 83
- 5.3 Intensifying and Extending the KCT Variable 87

5.4 Control Lines for Identity Plots 94
5.5 Connecting Two Tests 96
5.6 Building Item Banks 98
5.7 Banking the KCTB Data 106
5.8 Common Person Equating with the KCTB 109
5.9 Common Item Equating with the KCTB 112
5.10 Criterion Referencing the KCT Variable 118
5.11 Item Calibration Quality Control 121
5.12 Norm Referencing the KCT Variable 126

6 DESIGNING TESTS 129
6.1 Introduction 129
6.2 The Measurement Target 129
6.3 The Measuring Test 131
6.4 The Shape of a Best Test 133
6.5 The Precision of a Best Test 134
6.6 The Error Coefficient 135
6.7 The Design of a Best Test 137
6.8 The Complete Rules for Best Test Design 139

7 MAKING MEASURES 141
7.1 Using a Variable to Make Measures 141
7.2 Converting Scores to Measures by UCON, PROX and UFORM 142
7.3 Measures from Best Tests by UFORM 145
7.4 Individualized Testing 151
7.5 Status Tailoring 153
7.6 Performance Tailoring 156
7.7 Self-Tailoring 161
7.8 Person Fit and Quality Control 165
7.9 Diagnosing Misfit 170
7.10 Correcting a Measure 181

8 CHOOSING A SCALE 191
8.1 Introduction 191
8.2 Formulas for Making New Scales 192
8.3 The Least Measurable Difference 192
8.4 Defining the Spacing Factor 195
8.5 Normative Scaling Units: NITS 198
8.6 Substantive Scaling Units: SITS 199
8.7 Response Probability Scaling Units: CHIPS 201
8.8 Reporting Forms 205

APPENDICES 211
Table A 212 & 213
Table B 214 & 215
Table C 216

REFERENCES 217

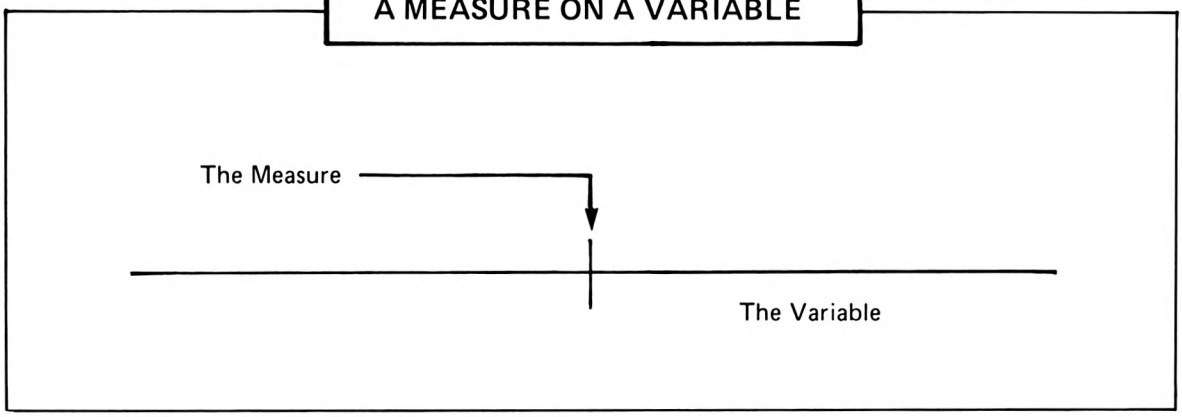
INDEX 220

1 THE MEASUREMENT MODEL

1.1 HOW TESTS ARE USED TO MEASURE

This book is about how to make and use mental tests. In order to do this successfully we must have a method for turning observations of test performance into measures of mental ability. The idea of a measure requires an idea of a variable on which the measure is located. If the variable is visualized as a line, then the measure can be pictured as a point on that line. This relationship between a measure and its variable is pictured in Figure 1.1.1.

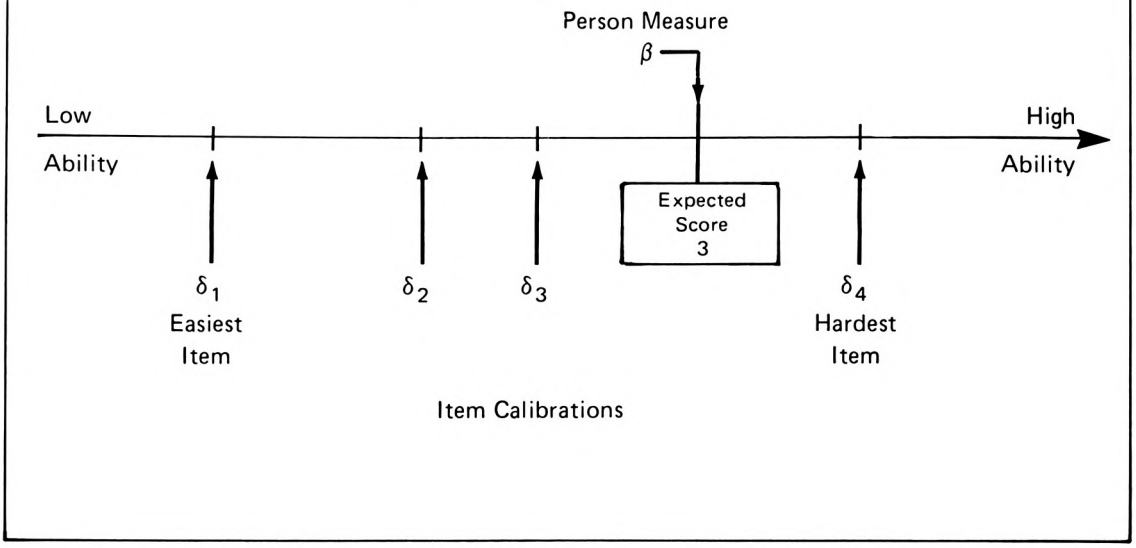
FIGURE 1.1.1
A MEASURE ON A VARIABLE



When we test a person, our purpose is to estimate their location on the line implied by the test. Before we can do this we must construct a test that defines a line. We must also have a way to turn the person's test performance into a location on that line. This book shows how to use test items to define lines and how to use responses to these items to position persons on these lines.

In order for a test to define a variable of mental ability, the items out of which the test is made must share a line of inquiry. This common line and its direction towards increasing ability can be pictured as an arrow with high ability to the right and low ability to the left. The meaning of this arrow is given by the test items which define it. If we use the symbols $\delta_1, \delta_2 \ldots \delta_i \ldots$, to represent the difficulty levels of items, then each δ_i marks the location of an item on the line. The δ's are the calibrations of the items along the variable and these calibrated items are the operational definition of what the variable measures. Hard items which challenge the most able persons define the high, or right, end of the line. Easy items which even the least able persons can usually do successfully define the low, or left, end of the line. Figure 1.1.2 shows a variable defined by four items spread across its length.

FIGURE 1.1.2
DEFINING A VARIABLE



A variable begins as a general idea of what we want to measure. This general idea is given substance by writing test items aimed at eliciting signs of the intended variable in the behavior of the persons. These test items become the operational definition of the variable. The intuition of the test builder and the careful construction of promising test items, however, are not enough. We must also gather evidence that a variable is in fact realized by the test items. We must give the items to suitable persons and analyze the resulting response patterns to see if the items fit together in such a way that responses to them define a variable.

In order to locate a person on this variable we must test them with some of the items which define the variable and then determine whether their responses add up to a position on the line. If we use the symbol β to represent the ability level of the person, then β marks their location on the line.

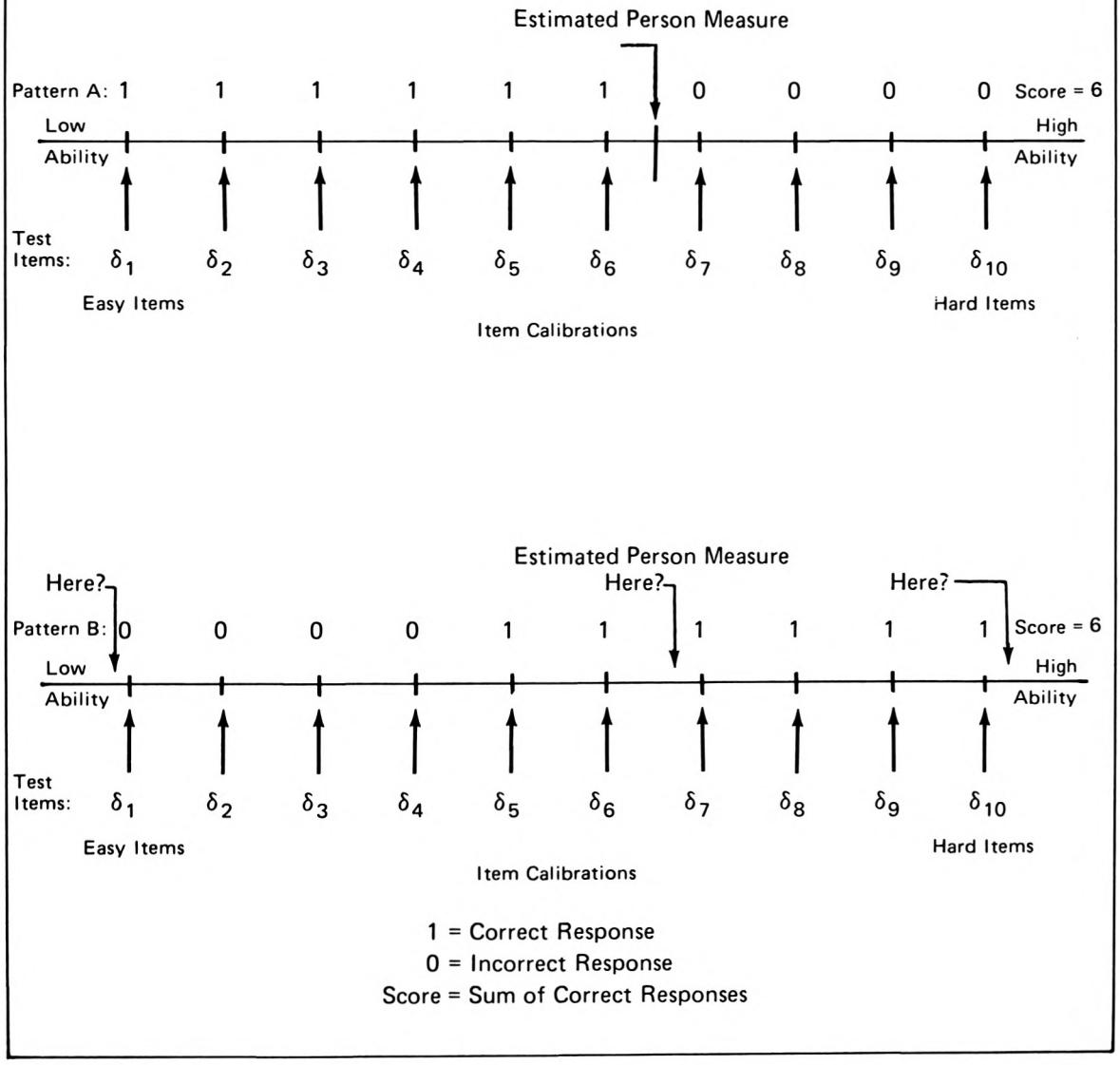
The person measure β shown in Figure 1.1.2 locates this person above the three easiest items and below the hardest one. Were this person to take a test made up of these four items, their most probable test score would be three and we would expect them to get the three easiest items correct and the fourth, hardest item, incorrect. This observation is more important than it might seem because it is the basis of all our methods for estimating person measures from test scores. When we want to know where a person is located on a variable, we obtain their responses to some of the items which define the variable. The only reasonable place to estimate their location from these data is in the region where their responses shift from mostly correct on easier items to mostly incorrect on harder ones.

Before we can estimate a person's measure from their score, however, we must examine their pattern of responses. We must see if their pattern is consistent with how we expect their items to elicit responses. When the items with which a person is tested have been calibrated along a variable from easy to hard, then we expect the person's response pattern to be more or less consistent with the difficulty order of these items along the

variable. We expect the person to succeed on items that ought to be easy for them and to fail on items that ought to be hard for them.

Figure 1.1.3 shows two response patterns to the same ten item test. The ten items are located along the variable at their levels of difficulty. Each pattern of responses is recorded above the line of the variable. The 1's represent correct answers. The 0's represent incorrect answers. Both patterns produce a score of six.

FIGURE 1.1.3
VALIDATING RESPONSE PATTERNS



In Pattern A the six easiest items are correct and the four hardest ones incorrect. It seems inconceivable to locate this person anywhere except in the region above δ_6, the most difficult item they get correct, but below δ_7, the least of the even more difficult items they get incorrect.

Pattern B, however, is very difficult to reconcile with the implications of a score of six. This person gets the six hardest items correct and the four easiest ones incorrect! If we try to locate this person above δ_{10}, the hardest items they get correct, we have to explain how they got the four easiest items incorrect. Could anyone be that careless? If, on the other hand, we try to locate them below δ_1, the easiest item they get incorrect, then how do we explain their getting the six hardest items incorrect? Every other location along the variable, such as between δ_6 and δ_7 for a score of six, is equally unsatisfactory as a "measure" for the person who produced Pattern B. This pattern of responses is not consistent with any location on the variable defined by these items. We are forced to conclude that something is wrong. Either the items used are miscalibrated or this person did not take them in the way we intended. In any case, no reasonable measure can be derived from Pattern B.

The Pattern B example is an important one because it shows us that even when we have constructed items that can define a valid variable we still have also to validate every person's response pattern before proceeding to use their score as a basis for estimating their measure. When item calibrations have been validated by enough suitable persons, then most of the response patterns we encounter among suitable persons will approximate Pattern A. However, the possibility of occurrences verging on Pattern B forces us to examine and validate routinely the response pattern of every person tested before we can presume to estimate a measure from their test score.

Four steps must be taken to use a test to measure a person. First, we must work out a clear idea of the variable we intend to make measures on. Second, we must construct items which are believable realizations of this idea and which can elicit signs of it in the behavior of the persons we want to measure. Third, we must demonstrate that these items when taken by suitable persons can lead to results that are consistent with our intentions. Finally, before we can use any person's score as a basis for their measure, we must determine whether or not their particular pattern of responses is, in fact, consistent with our expectations.

1.2 HOW SCORES ARE USED

A test score is intended to locate a person on the variable defined by the test items taken. Nearly everyone who uses test scores supposes that the person's location on the variable is satisfactorily determined either by the score itself, or by some linear function of the score such as a percent correct or a norm-based scale value. It is taken for granted that the score, or its scale equivalent, tells us something about the person tested that goes beyond the moment or materials of the testing. It is also taken for granted that scores are suitable for use in the arithmetic necessary to study growth and compare groups. But do scores actually have the properties necessary to make it reasonable to use them in these ways?

In order for a particular score to have meaning it must come from a response pattern which is consistent with items that define a variable. But even the demonstration of item validity and response validity does not guarantee that the score will be useful. In order to generalize about the person beyond their score, in order to discover what their score implies, we must also take into account and adjust for the particulars of the test items used. How, then, does a person's test score depend on the characteristics of the items in the test they take?

FIGURE 1.2.1

HOW SCORES DEPEND ON THE LEVEL AND SPREAD OF TEST ITEM DIFFICULTIES

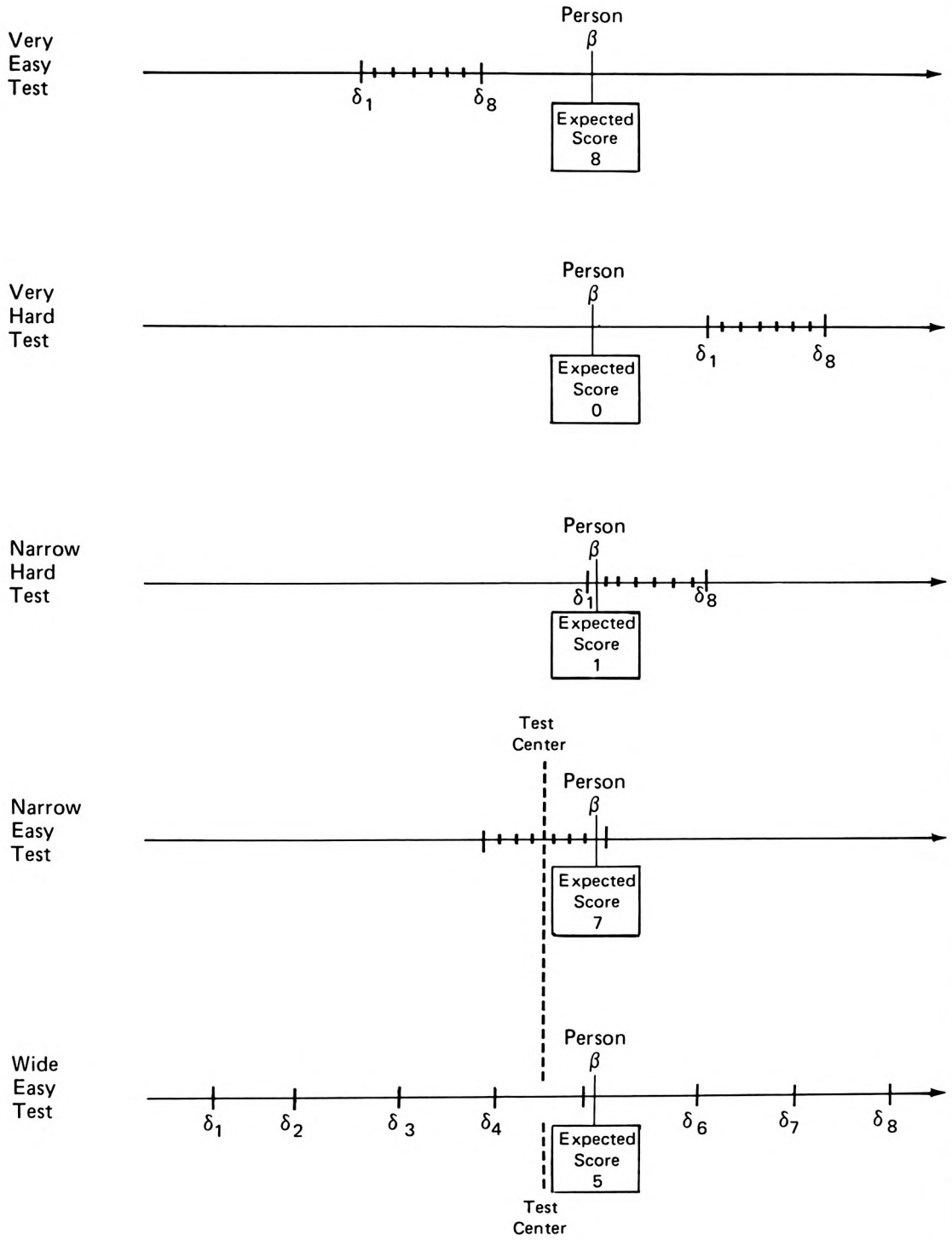


Figure 1.2.1 shows what can happen when one person at a particular ability level takes five different tests all of which measure on the same variable but which differ in the level and spread of their item difficulties from easy to hard and narrow to wide. The difficulties of the eight items in each test are marked on the line of the variable. In order to see each test separately we have redrawn the line of the variable five times, once for each test.

The ability of the person on the measure is also marked on each line so that we can see how this person stands with respect to each test. While each test has a different position on the variable depending on the difficulties of its items, this person's position, of course, is the same on each line. Figure 1.2.1 also shows the scores we would expect this person most often to get on these five tests.

The first, Very Easy Test, has items so easy for this person that we expect a test score of eight. The second, Very Hard Test, has such hard items that we expect a score of zero. The third, Narrow Hard Test, has seven of its items above the person's ability and one below. In this situation the score we would expect most often to see would be a one. The fourth, Narrow Easy Test, has seven of its items below the person's ability and so we expect a score of seven. Finally the fifth, Wide Easy Test, has five items which should be easy for them. Even though this test is centered at the same position on the variable as the Narrow Easy Test just above it in Figure 1.2.1 and so has the same average difficulty level, nevertheless, because of its greater width in item difficulty, we expect only a score of five.

For one person we have five expected scores: zero, one, five, seven and eight! Although we know the person's ability does not change, the five different scores, as they stand, suggest five different abilities. Test scores obviously depend as much on the item characteristics of the test as on the ability of the person taking the test.

If the meaning of a test score depends on the characteristics of the test items, however, then before we can determine a person's ability from their test score we must "adjust" their score for the effects of the particular test items from which that particular score comes. This adjustment must be able to turn test-bound scores into measures of person ability which are test-free.

Unfortunately, with test scores like zero, in which there is no instance of success, and the eight of our example, in which there is no instance of failure, there is no satisfactory way to settle on a finite measure for the person. All we can do in those situations is to observe that the person who scored all incorrect or all correct is substantially below or above the operating level of the test they have taken. If we wish to estimate a finite measure for such a person, then we will have to find a test for them which is more appropriate to their level of ability.

We might be tempted to interpret perfect scores as "complete mastery." But unless the test in question actually contained the most difficult items that could ever be written for this variable there would always be the possibility of other items which were even more difficult. These more difficult items might produce incorrect answers, even with our perfectly scoring person, revealing that mastery was not complete after all. When a test is extremely easy, of course, everyone recognizes that even a perfect score is quite consistent with intermediate ability.

The dependence of test scores on item difficulty is a problem with which most test users are familiar. Almost everyone realizes that fifty percent correct on an easy test does not mean as much as fifty percent correct on a hard test. Some test users even realize that seventy-five percent correct on a narrow test does not imply as much ability as seventy-five percent correct on a wide test. But there is another problem in the use of test scores which is often overlooked.

It is common practice to compute differences in test scores to measure growth, to combine test scores by addition and subtraction in order to compare groups and to add and subtract squares and cross-products of test scores in order to do regression analysis. But when these simple arithmetic operations are applied to test scores the results are always slightly distorted and can be substantially misleading. Although test scores usually estimate the order of persons' abilities rather well, they never estimate the spacing satisfactorily. Test scores are not linear in the measures they imply and for which they are used.

In the statistical use of test scores, floor and ceiling effects are occasionally recognized. But they are almost never adjusted for. These boundary effects cause any fixed differences of score points to vary in meaning over the score range of the test. The distance on the variable a particular difference in score points implies is not the same from one end of the test to the other. A difference of five score points, for example, implies a larger change in ability at the ends of a test than in the middle.

Figure 1.2.2 illustrates this problem with test scores. We show two persons with measures, β_A and β_B, who are a fixed distance apart on the same variable. Both persons are administered five different tests all measuring on this variable. The persons' locations and hence their measurable difference on the variable remain the same from test to test, but their most probable scores vary widely. This is because the five tests differ in their item difficulty level, spread and spacing. Let's see how the resulting expected scores reflect the fixed difference between these two persons.

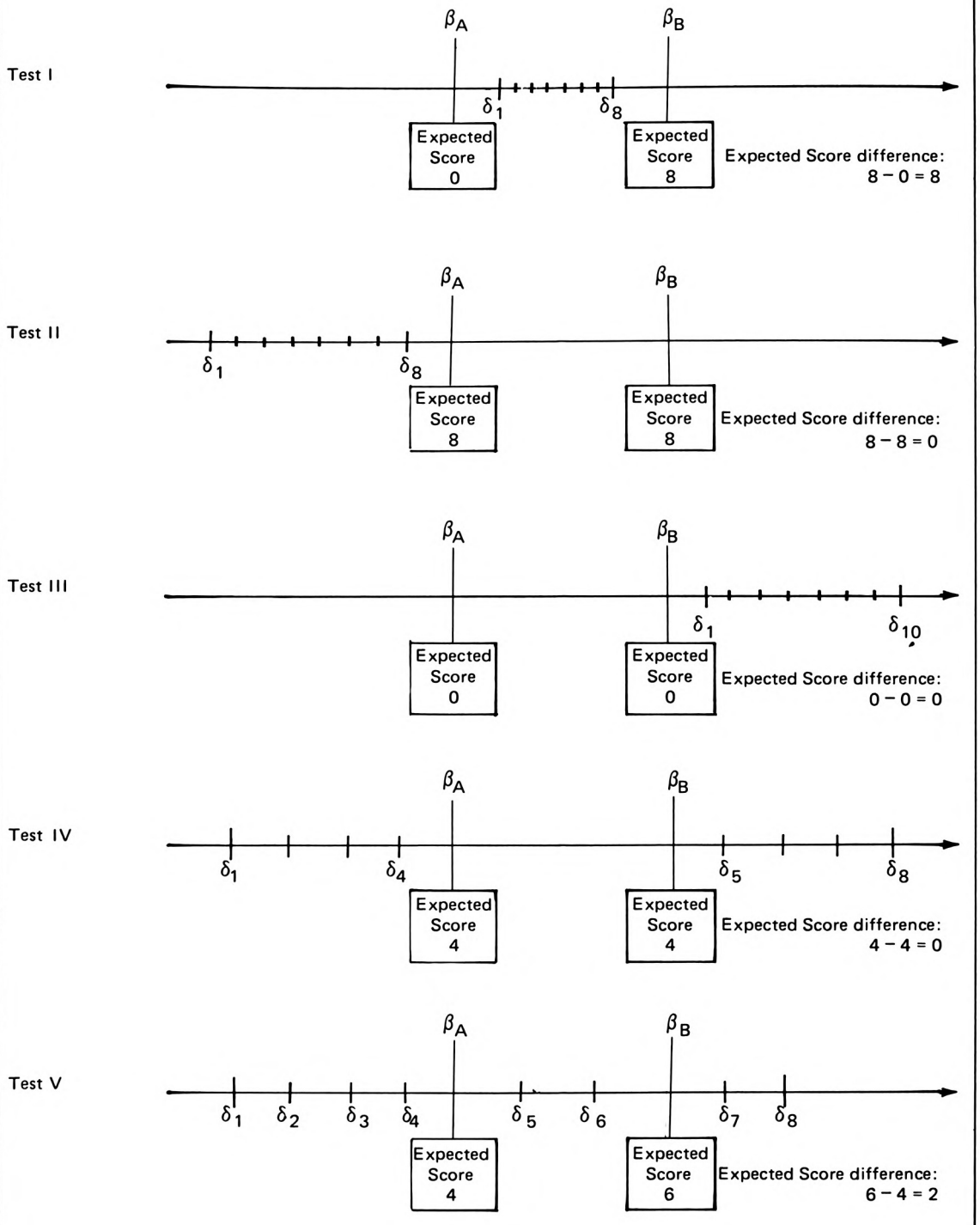
Test I is composed of eight items all of which fall well between Person A and Person B. We expect Person A to get none of these items correct for a score of zero while we expect Person B to get all eight items correct for a score of eight. On this test their abilities will usually appear to be eight score points apart. That is as far apart in ability as it is possible to be on this test.

Test II is composed of eight items all of which are well below both persons. We expect both persons to get scores of eight because this test is too easy for both of them. Now their expected score difference in test scores is zero and their abilities will usually appear to be the same!

Test III is composed of eight very hard items. Now we expect both persons to get scores of zero because this test is too hard for them. Once again their expected score difference is zero and their abilities will usually appear to be the same.

Test I was successful in separating Persons A and B. Tests II and III failed because they were too far off target. Perhaps it is only necessary to center a test properly in order to observe the difference between two persons.

FIGURE 1.2.2
THE NONLINEARITY OF SCORES



Test IV is centered between Person A and Person B but its items are so spread out that there is a wide gap in its middle into which Person A and Person B both fall. The result is that both persons can be expected to achieve scores of four because four items are too easy and four items are too hard for both of them. Even for this test which is more or less centered on their positions, their expected score difference is zero and their abilities will still usually appear to be the same.

Test V, at last, is both wide and fairly well centered on Persons A and B. It contains two items which fall between their positions and therefore separate them. We expect Person A to get the four easiest items correct for a most probable score of four. As for Person B, however, we expect them not only to get the same four items correct but also the next two harder ones because these two items are also below Person B's ability level. Thus on Test V the expected difference in scores between Person's A and B becomes two. On this test their abilities will usually appear to be somewhat, but not extremely, different.

What can we infer about the differences in ability between Persons A and B from scores like these? Persons A and B will tend to appear equally able on Tests II, III, and IV, somewhat different on Test V and as different as possible on Test I. If differences between the test scores of the same two persons can be made to vary so widely merely by changing the difficulties of the items in the test, then how can we use differences in test scores to study ability differences on a variable?

The answer is, we can't. Not as they stand. In order to use test scores, which are not linear in the variable they imply, to analyze differences we must find a way to transform the test scores into measures which approximate linearity.

Test scores always contain a potentially misleading distortion. If we intend to use test results to study growth and to compare groups, then we must use a method for making measures from test scores which marks locations along the variable in an equal interval or linear way.

In this section we have illustrated two serious problems with test scores. The first illustration shows how test scores are test-bound and how we have to adjust them for the characteristics of their test items before we can use the scores as a basis for measurement. The second illustration shows how test scores do not mark locations on their variable in a linear way and how we need to transform test scores into measures that are linear before we can use them to study growth or to compare groups.

1.3 WHAT HAPPENS WHEN A PERSON TAKES AN ITEM

The discussions in Sections 1.1 and 1.2 establish our need for 1) valid items which can be demonstrated to define a variable, 2) valid response patterns which can be used to locate persons on this variable, 3) test-free measures that can be used to characterize persons in a general way and 4) linear measures that can be used to study growth and compare groups. Now we must build a method that comes to grips with these requirements.

The responses of individual persons to individual items are the raw data with which we begin. The method we develop must take these data and make from them item calibrations and person measures with the properties we require. Figure 1.3.1 shows a very

simple data matrix containing the responses of eight persons to a five item test. The five items are named at the top of the matrix. The eight persons are named at the left. The response of each person to each item is indicated by "1" for a correct response and "0" for an incorrect response. Notice that the responses in Figure 1.3.1 have been summed across the items and entered on the right side of the matrix as person scores and down the persons and entered at the bottom of the matrix as item scores.

FIGURE 1.3.1
A DATA MATRIX
OF OBSERVED RESPONSES

Person Name	Item Name 1	2	3	4	5	Person Score
a	1	0	0	0	0	1
b	0	1	0	0	0	1
c	1	1	0	0	0	2
d	1	0	1	0	0	2
e	1	1	1	0	0	3
f	1	1	0	1	0	3
g	1	1	1	1	0	4
h	1	1	1	0	1	4
	7	6	4	2	1	20
			Item Score			

Figure 1.3.1 shows what the basic data look like. But before we can put these data to work we must answer a fundamental question. Where do we think these data come from? What are these item and person scores supposed to tell us about items and persons? How do we think these patterns of 1's and 0's are produced? In order to figure out how to use these data we must set up a reasonable model for what we suppose happens when a person attempts to answer an item.

We would like a person v's ability β_v, that is their location on the variable, to govern how far along the variable we can expect them to produce correct responses to items. Indeed that is the only situation in which we can use item difficulties and a person's responses to them as the basis for measuring the person.

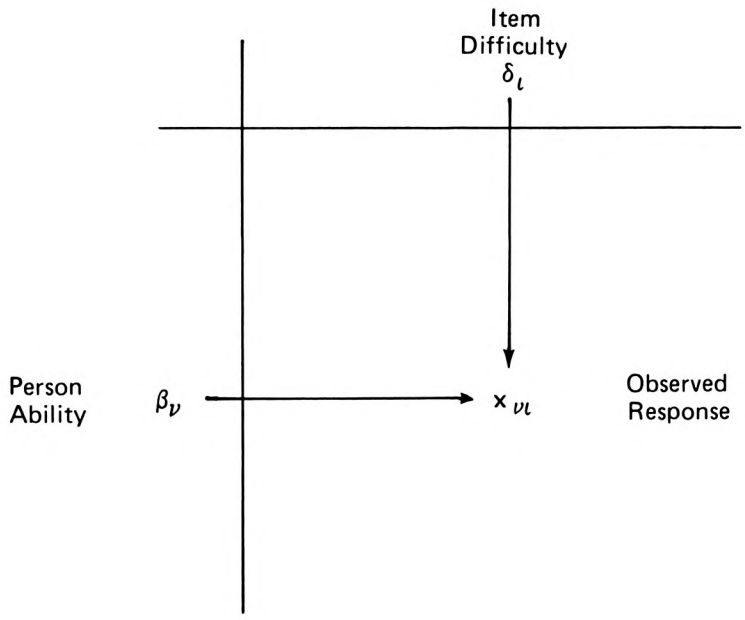
Of course we can think of other factors which might affect a person's responses. If items are multiple-choice, some guessing is bound to occur and persons differ in how much guessing they are willing to engage in. The possibilities of disturbing influences which interfere with the clear expression and hence the unambiguous observation of ability are endless. But, if it is really the person's ability that we hope to measure, then

it would be unreasonable not to do our best to arrange things so that it is the person's ability which dominates their test behavior. Indeed, isn't that what good test administration practices are for, namely, to control and minimize the intrusion of interfering influences.

We would also like item ι's difficulty δ_ι, that is its location on the variable, to determine how far along the variable we can expect correct responses to that item to occur. As with persons, we can think up item characteristics, such as discrimination and vulnerability to guessing, which might modify persons' responses to them. Some psychometricians attempt to estimate these additional item characteristics even though there are good reasons to expect that all such attempts must, in principle, fail. But, again, it hardly seems reasonable not to do our best to arrange things so that it is an item's difficulty which dominates how persons of various abilities respond to that item. In any case, the fact is that whenever we use unweighted scores as our test results we are assuming that, for all practical purposes, it is item difficulties, and person abilities, that dominate person responses.

FIGURE 1.3.2

THE ESSENTIAL CONDITIONS CAUSING A RESPONSE



When the response is "correct",
$x = 1$

When the response is "incorrect",
$x = 0$

These considerations lead us to set up a response model that is the simplest representation possible. Figure 1.3.2 diagrams person ν with ability β_ν acting on item ι with difficulty δ_ι to produce the response $x_{\nu\iota}$. These are the essential elements we will take into account when we try to explain the data in Figure 1.3.1. Figure 1.3.2 proposes that the response $x_{\nu\iota}$ which occurs when person ν takes item ι can be thought of as governed by the person's ability β_ν and the item's difficulty δ_ι and nothing else.

Our next step is to decide how we want person ability β_ν and item difficulty δ_ι to interact in order to produce $x_{\nu\iota}$. What is a reasonable and useful way to set up a mathematical relation between β_ν and δ_ι? Since we require that β_ν and δ_ι represent locations along one common variable which they share, it is their difference $(\beta_\nu - \delta_\iota)$ which is the most convenient and natural formulation of their relation.

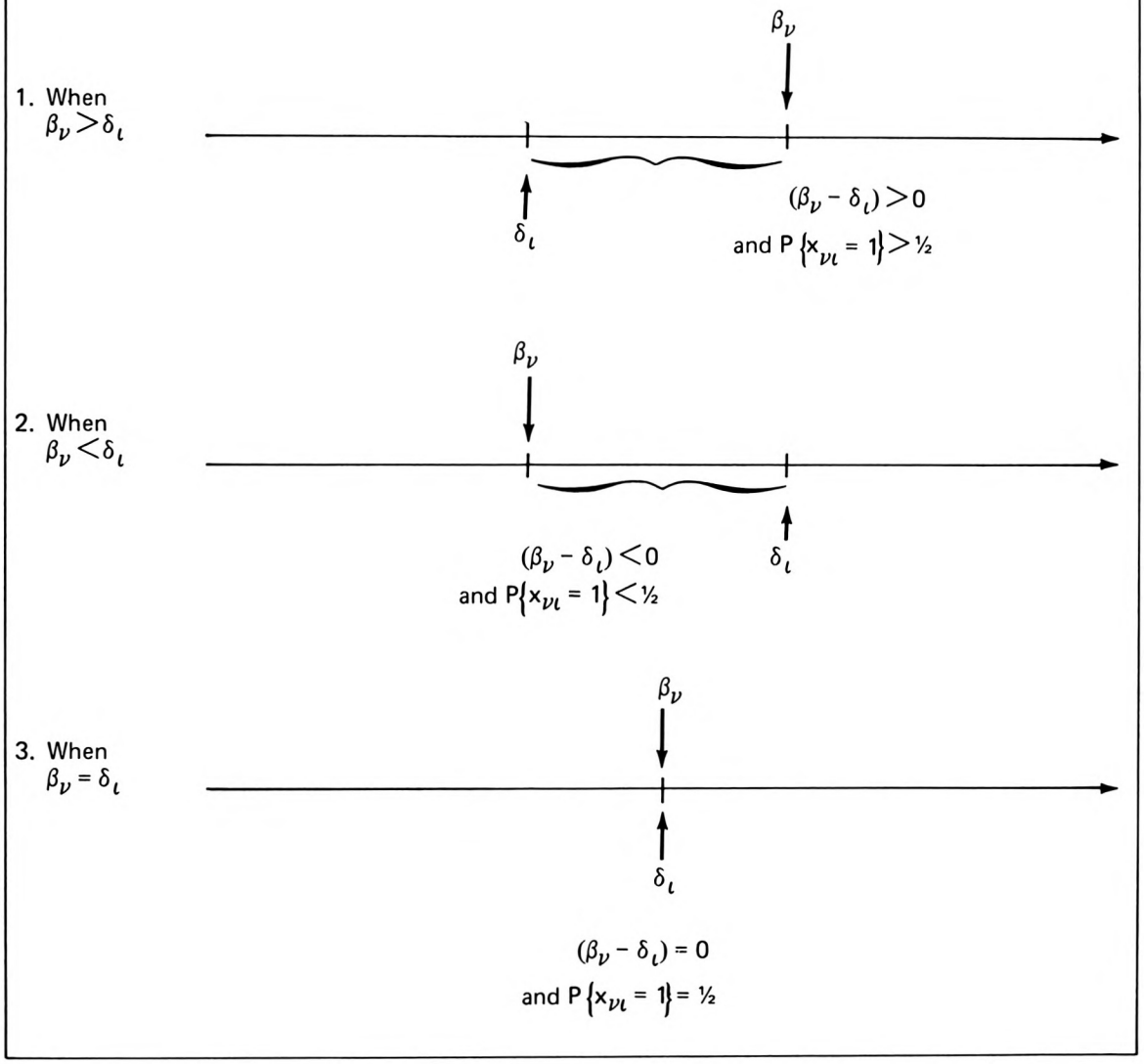
Identifying the difference $(\beta_\nu - \delta_\iota)$, however, does not finish our work because we must also decide how we want this difference to govern the value of the response $x_{\nu\iota}$. Even when a person is more able than an item is difficult, so that their β_ν is greater than the item's δ_ι, it will occasionally happen that this person nevertheless fails to give a correct answer to that relatively easy item so that the resulting value of $x_{\nu\iota}$ is "0". It will also happen occasionally that a person of moderate ability nevertheless succeeds on a very difficult item. Obviously it is going to be awkward to force a deterministic relationship onto the way $(\beta_\nu - \delta_\iota)$ governs the value of response $x_{\nu\iota}$. A better way to deal with this problem is to acknowledge that the way the difference $(\beta_\nu - \delta_\iota)$ influences the response $x_{\nu\iota}$ can only be probabilistic and to set up our response model accordingly.

Figure 1.3.3 shows how it would be most reasonable to have the difference $(\beta_\nu - \delta_\iota)$ affect the probability of a correct response. When β_ν is larger than δ_ι, so that the ability level of person ν is greater than the difficulty level of item ι and their difference $(\beta_\nu - \delta_\iota)$ is greater than zero, then we want the probability of a correct answer to be greater than one half. When, on the other hand, the ability level of person ν is less than the difficulty level of item ι, so that their difference $(\beta_\nu - \delta_\iota)$ is less than zero, then we want the probability of a correct answer to be less than one half. Finally, when the levels of person ability and item difficulty are the same so that their difference $(\beta_\nu - \delta_\iota)$ is zero, then the only probability that seems reasonable to assign to a correct (or to an incorrect) answer is exactly one half.

The curve in Figure 1.3.4 summarizes the implications of Figure 1.3.3 for all reasonable relationships between probabilities of correct responses and differences between person ability and item difficulty. This curve specifies the conditions our response model must fulfill. The differences $(\beta_\nu - \delta_\iota)$ could arise in two ways. They could arise from a variety of person abilities reacting to a single item or they could arise from a variety of item difficulties testing the ability of one person. When the curve is drawn with ability β as its variable so that it describes an item, it is called an item characteristic curve (ICC) because it shows the way the item elicits responses from persons of every ability. When the curve is drawn with difficulty δ as its variable so that it describes how a person responds to a variety of items, we can call it a person characteristic curve (PCC).

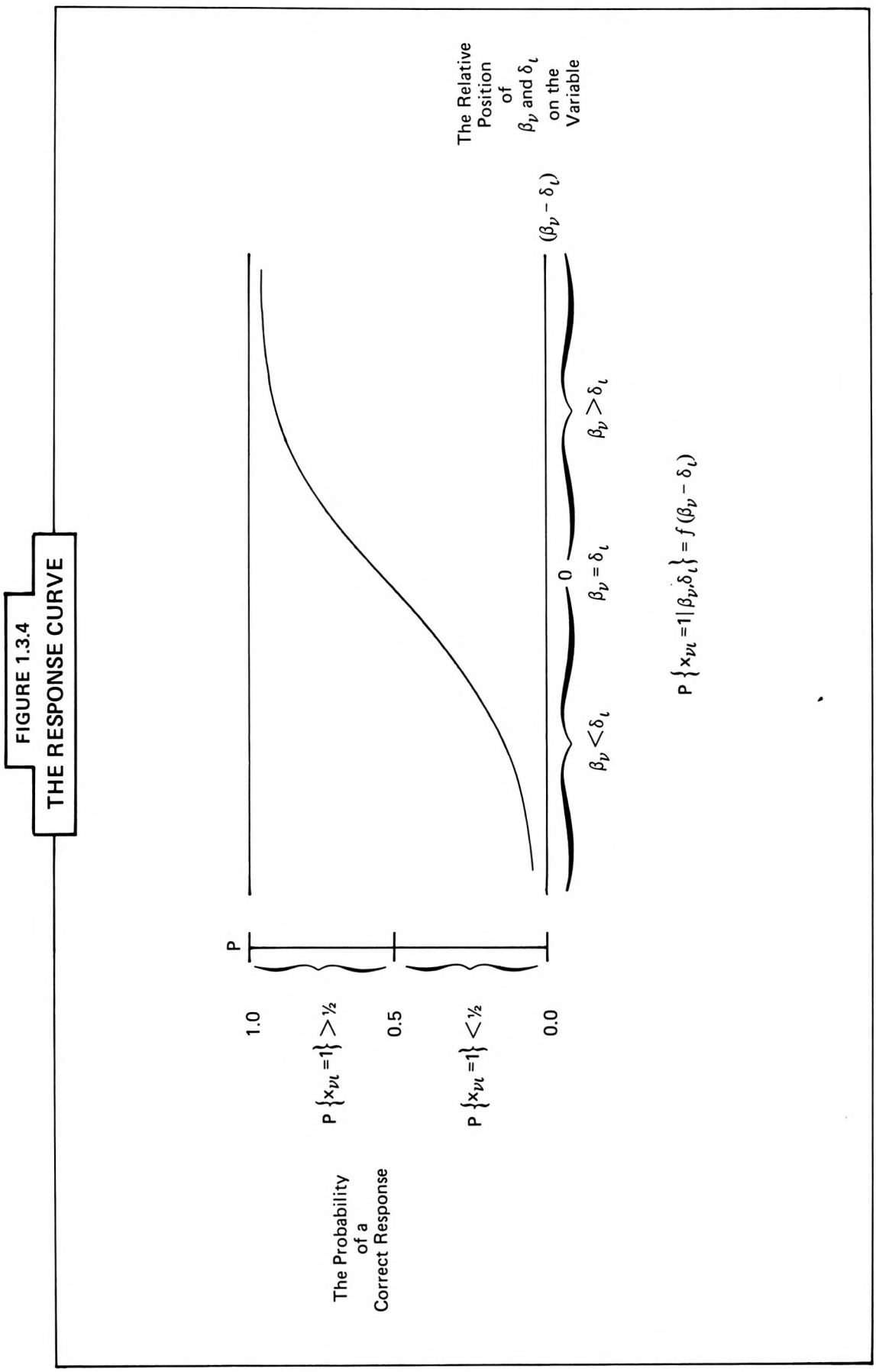
FIGURE 1.3.3

HOW DIFFERENCES BETWEEN PERSON ABILITY AND ITEM DIFFICULTY OUGHT TO AFFECT THE PROBABILITY OF A CORRECT RESPONSE



The curve in Figure 1.3.4 is a picture of the response model we require in order to solve the problem of how the parameters, β_v and δ_i which we want to estimate, depend on the data x_{vi} we can observe. To measure a person, we must estimate β_v and to calibrate an item we must estimate δ_i. In order to estimate either of these parameters from the observed responses of persons to items we must construct a mathematical formulation which is true to the relationship drawn in Figure 1.3.4 and which relates β_v, δ_i and x_{vi} in a useful way. This formulation must also be able to show us how to use data of the kind given in Figure 1.3.1 to make estimates of person ability which are test-free and estimates of item difficulty which are sample-free.

FIGURE 1.3.4
THE RESPONSE CURVE



1.4 THE RASCH MODEL

In order to construct a workable mathematical form for the curve in Figure 1.3.4 we begin by combining the parameters, β_v for person ability and δ_i for item difficulty through their difference $(\beta_v - \delta_i)$. We want this difference to govern the probability of what is supposed to happen when person v uses their ability β_v against the difficulty δ_i of item i. But the difference $(\beta_v - \delta_i)$ can vary from minus infinity to plus infinity while the probability of a successful response must remain between zero and one. To deal with this we apply the difference $(\beta_v - \delta_i)$ as an exponent of the natural constant e = 2.71828 . . . and write the result as

$$e^{(\beta_v - \delta_i)} = \exp(\beta_v - \delta_i).$$

This exponential expression varies between zero and plus infinity and we can bring it into the interval between zero and one by forming the ratio

$$\exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)].$$

This formulation has a shape which follows the ogive in Figure 1.3.4 quite well. It can be used to specify the probability of a successful response as

$$P\{x_{vi} = 1 | \beta_v, \delta_i\} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)] \qquad [1.4.1]$$

which is the Rasch model.

Any mathematical form which describes an ogive of the shape in Figure 1.3.4 could provide a solution to the linearity problem by transforming scores which are restricted between 0 and 100 percent into "measures" which run from minus infinity to plus infinity.

Any mathematical form which relates the probability of x_{vi} to the difference between β_v and δ_i and which has estimable parameters could allow us to study item and response validity. All we have to do is to specify a workable model for how $(\beta_v - \delta_i)$ governs the probability of x_{vi}, use this model to estimate β_v and δ_i from some data and then examine the way these data fit with predictions calculated from the model.

Any ogive and any formulation, however, will not do. In fact, only the formulation of Equation 1.4.1, the Rasch model, allows us to estimate β_v and δ_i independently of one another in such a way that the estimates β_v are freed from the effects of the δ_i and the estimates d_i are freed from the effects of the β_v's.

The logistic function in Equation 1.4.1 provides a simple, useful response model that makes both linearity of scale and generality of measure possible. Although biometricians have used the logistic function since 1920, it was the Danish mathematician Georg Rasch (1960) who first appreciated its psychometric significance. Rasch calls the special characteristic of the simple logistic function which makes generality in measurement possible "specific objectivity." He and others have shown that there is no alternative mathematical formulation for the ogive in Figure 1.3.4 that allows estimation of the person measures β_v and the item calibrations δ_i independently of one another (Rasch, 1961, 1967; Andersen, 1973, 1977; Barndorff-Nielsen, 1978). When the estimators for β_v and δ_i are derived by maximizing a conditional liklihood they are unbiased, consistent,

efficient, and sufficient (Andersen, 1970, 1971, 1972a, 1973, 1977; Haberman, 1977). Simple approximations for these conditional maximum likelihood estimators which are accurate enough for almost all practical purposes are described in Wright and Panchapakesan (1969), Wright and Douglas (1975a, 1975b, 1977a, 1977b) and Wright and Mead (1976). These procedures have been useful in a wide variety of applications (Connolly, Nachtman and Pritchett, 1971; Woodcock, 1974; Willmott and Fowles, 1974; Rentz and Bashaw, 1975, 1977; Andrich, 1975; Mead, 1975; Wright and Mead, 1977; Cornish and Wines, 1977; Draba, 1978; Elliott, Murray and Pearson, 1977.

TABLE 1.4.1

PERSON ABILITY AND ITEM DIFFICULTY IN LOGITS AND THE RASCH PROBABILITY OF A RIGHT ANSWER

Person Ability	Item Difficulty	Difference	Right Answer		Information in a Response
			Odds	Probability	
β_v	δ_i	$(\beta_v - \delta_i)$	$\exp(\beta_v - \delta_i)$	π_{vi}	I_{vi}
5	0	5	148.	.99	.01
4	0	4	54.6	.98	.02
3	0	3	20.1	.95	.05
2	0	2	7.39	.88	.11
1	0	1	2.72	.73	.20
0	0	0	1.00	.50	.25
0	1	- 1	0.368	.27	.20
0	2	- 2	0.135	.12	.11
0	3	- 3	0.050	.05	.05
0	4	- 4	0.018	.02	.02
0	5	- 5	0.007	.01	.01

$$\pi_{vi} = \exp(\beta_v - \delta_i) / [1 + \exp(\beta_v - \delta_i)]$$

$$I_{vi} = \pi_{vi}(1 - \pi_{vi})$$

We can see in Equation 1.4.1 that when person v is smarter than item i is difficult, then β_v is more than δ_i, their difference is positive and the probability of success on item i is greater than one half. The more the person's ability surpasses the item's difficulty, the greater this positive difference and the nearer the probability of success comes to one. But when the item is too hard for the person, then β_v is less than δ_i, their difference is negative and the person's probability of success is less than one half. The more the item overwhelms the person, the greater this negative difference becomes and the nearer the probability of success comes to zero.

The mathematical units for β_v and δ_i defined by this model are called "logits." A person's ability in logits is their natural log odds for succeeding on items of the kind chosen to define the "zero" point on the scale. And an item's difficulty in logits is its natural log odds for eliciting failure from persons with "zero" ability.

Table 1.4.1 gives examples of various person abilities and item difficulties in logits, their differences $(\beta_v - \delta_i)$ and the success probabilities which result. The first six rows illustrate various person abilities and their success probabilities when provoked by items of zero difficulty. The last six rows give examples of various item difficulties and the probabilities of success on them by persons with zero ability.

The origin and scale of the logits used in Table 1.4.1 are arbitrary. We can add any constant to all abilities and all difficulties without changing the difference $(\beta_v - \delta_i)$. This means that we can place the zero point on the scale so that negative difficulties and abilities do not occur. We can also introduce any scaling factor we find convenient including one large enough to eliminate any need for decimal fractions. Chapter 8 investigates these possibilities in detail.

The last column of Table 1.4.1 gives the relative information $I_{vi} = \pi_{vi}(1 - \pi_{vi})$ available in a response observed at each $(\beta_v - \delta_i)$. When item difficulty δ_i is within a logit of person ability β_v, the information about either δ_i or β_v in one observation is greater than .20. But when item difficulty is more than two logits off target, the information is less and .11 and for $|\beta_v - \delta_i| > 3$ less than .05. The implications for efficient calibration sampling and best test design are that responses in the $|\beta_v - \delta_i| < 1$ region are worth more than twice as much for calibrating items or measuring persons as those outside of $|\beta_v - \delta_i| > 2$ and more than four times as much as those outside of $|\beta_v - \delta_i| > 3$.

1.5 USING THE RASCH MODEL FOR CALIBRATING AND MEASURING

We have established the need for an explicit approach to measurement and shown how measurement problems can be addressed with a model for what happens when a person takes an item. Now we are ready to work through the mathematics of this model in order to find out how we can use the model to calibrate items and measure persons. The model specifies the probability of person v with ability β_v giving response x_{vi} to item i with difficulty δ_i as

$$P\{x_{vi} \mid \beta_v, \delta_i\} = \exp[x_{vi}(\beta_v - \delta_i)] / [1 + \exp(\beta_v - \delta_i)] \qquad [1.5.1]$$

The response x_{vi} takes only two values,

$x_{vi} = 0$ when the response is incorrect and
$x_{vi} = 1$ when the response is correct.

When we insert each of these values of x_{vi} into Equation 1.5.1 we find that it breaks down into the complementary expressions

$$P\{x_{vi} = 1 \mid \beta_v, \delta_i\} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)] \qquad [1.5.2]$$

for a correct response and

$$P\{x_{vi} = 0 \mid \beta_v, \delta_i\} = 1/[1 + \exp(\beta_v - \delta_i)] \qquad [1.5.3]$$

for an incorrect response. These two expressions add up to one because together they cover everything that can happen to $x_{\nu\iota}$.

When a set of L items is administered to a sample of N persons, the result is an N by L collection of responses $((x_{\nu\iota}))$ for which $\nu = 1, N$ and $\iota = 1, L$ and the double parentheses (( )) remind us that a whole N by L table of $x_{\nu\iota}$'s is implied. These data can be displayed in a matrix like the one shown in Figure 1.5.1. The marginal sums of the rows and colums of this data matrix are the person scores

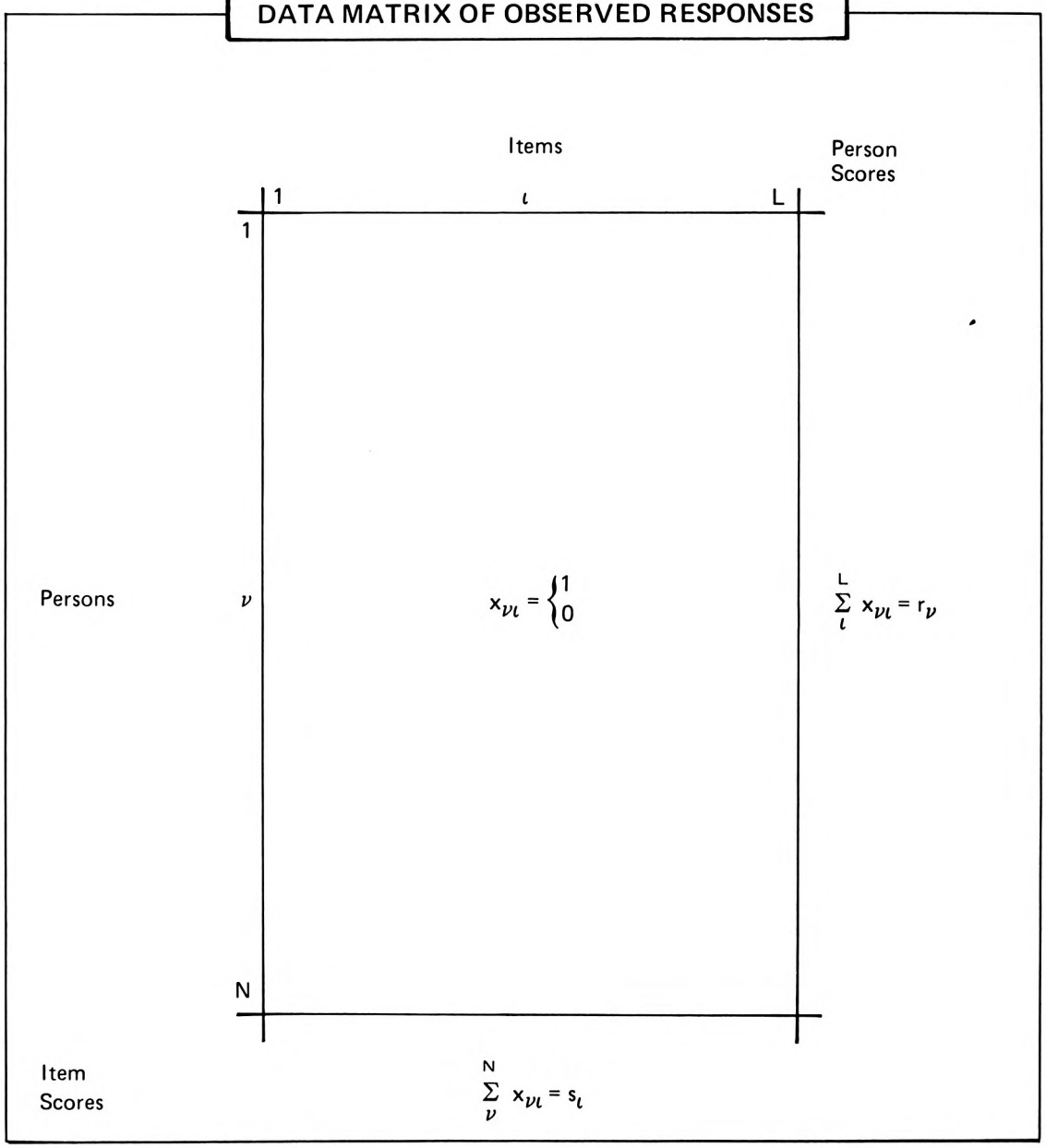
$$r_\nu = \sum_{\iota}^{L} x_{\nu\iota} \qquad \nu = 1,N$$

and the item scores

$$s_\iota = \sum_{\nu}^{N} x_{\nu\iota} \quad . \qquad \iota = 1,L$$

FIGURE 1.5.1

DATA MATRIX OF OBSERVED RESPONSES



What happens when we analyze these data as though they were governed by the model of Equation 1.5.1? According to that model the only systematic influences on the production of the $x_{v\iota}$'s are the N person abilities (β_v) and the L item difficulties (δ_ι). As a result, apart from these parameters, the $x_{v\iota}$'s are modeled to be quite independent of one another. This means that the probability of the whole data matrix ($(x_{v\iota})$), given the model and its parameters (β_v) and (δ_ι), can be expressed as the product of the probabilities of each separate $x_{v\iota}$ given by Equation 1.5.1 continued over all $v = 1, N$ and all $\iota = 1, L$.

This continued product is

$$P\{(x_{v\iota})|(\beta_v),(\delta_\iota)\} = \prod_{v}^{N}\prod_{\iota}^{L}\left\{\frac{\exp[x_{v\iota}(\beta_v - \delta_\iota)]}{1 + \exp(\beta_v - \delta_\iota)}\right\}. \qquad [1.5.4]$$

When we move the continued product operators $\prod_{v}^{N}$ and $\prod_{\iota}^{L}$ in the numerator of Equation 1.5.4 into the exponential expression

$$\exp[x_{v\iota}(\beta_v - \delta_\iota)],$$

they become the summation operators

$$\sum_{v}^{N} \text{ and } \sum_{\iota}^{L} \text{ so that}$$

$$\prod_{v}^{N}\prod_{\iota}^{L} \exp[x_{v\iota}(\beta_v - \delta_\iota)] = \exp\left[\sum_{v}^{N}\sum_{\iota}^{L} x_{v\iota}(\beta_v - \delta_\iota)\right].$$

Then, since

$$\sum_{v}^{N}\sum_{\iota}^{L} x_{v\iota}\beta_v = \sum_{v}^{N} r_v\beta_v$$

and

$$\sum_{v}^{N}\sum_{\iota}^{L} x_{v\iota}\delta_\iota = \sum_{\iota}^{L} s_\iota\delta_\iota,$$

Equation 1.5.4 becomes

$$P\{(x_{v\iota})|(\beta_v),(\delta_\iota)\} = \frac{\exp\left[\sum_{v}^{N} r_v\beta_v - \sum_{\iota}^{L} s_\iota\delta_\iota\right]}{\prod_{v}^{N}\prod_{\iota}^{L}[1 + (\exp(\beta_v - \delta_\iota)]} \qquad [1.5.5]$$

Equation 1.5.5 is important because it shows that in order to estimate the parameters (β_v) and (δ_ι), we need only the marginal sums of the data matrix, (r_v) and (s_ι). This is because that is the only way the data ($(x_{v\iota})$) appear in Equation 1.5.5. Thus the person scores (r_v) and item scores (s_ι) contain all the modelled information about person measures and item calibrations.

Finally, the numerator of Equation 1.5.5 can be factored into two parts so that the model probability of the data matrix becomes

$$P\{((x_{vi}))|(\beta_v),(\delta_i)\} = \frac{\left[\exp\left(\sum_v^N r_v \beta_v\right)\right]\left[\exp\left(-\sum_i^L s_i \delta_i\right)\right]}{\prod_v^N \prod_i^L [1 + \exp(\beta_v - \delta_i)]} \qquad [1.5.6]$$

Equation 1.5.6 is important because it shows that the person and item parameters can be estimated independently of one another. The separation of

$$\left(\sum_v^N r_v \beta_v\right)$$

and

$$\left(\sum_i^L s_i \delta_i\right)$$

in Equation 1.5.6 makes it possible to condition either set of parameters out of Equation 1.5.6 when estimating the other set. This means, in the language of statistics, that the scores (r_v) and (s_i) are sufficient for estimating the person measures and the item calibrations.

Because of this we can use the person scores (r_v) to remove the person parameters (β_v) from Equation 1.5.6 when calibrating items. This frees the item calibrations from the modelled characteristics of the persons and in this way produces sample-free item calibrations. As for measuring persons, we could use the item scores (s_i) to remove the item parameters (δ_i) from Equation 1.5.6. When we come to person measurement, however, we will find it more convenient to work directly from the estimated item calibrations (d_i).

There are several ways that Equation 1.5.6 can be used to estimate values for β_v and δ_i. The ideal way is to use the sufficient statistics for persons (r_v) to condition person parameters (β_v) out of the equation. This leaves a conditional liklihood involving only the item parameters (δ_i) and they can be estimated from this conditional liklihood (Fischer and Scheiblechner, 1970; Andersen, 1972a, 1972b; Wright and Douglas, 1975b, 1977b; Allerup and Sorber, 1977; Gustafsson, 1977).

But this ideal method is impractical and unnecessary. Computing times are excessive. Round-off errors limit application to tests of fifty items at most. And, in any case, results are numerically equivalent to those of quicker and more robust methods. A convenient and practical alternative is to use Equation 1.5.6 as it stands. To learn more about this unconditional estimation of item parameters see Wright and Panchapakesan (1969), Wright and Douglas (1975b, 1977a, 1977b) and Chapter 3, Section 3.4. of this book.

Even this unconditional method, however, is often unnecessarily detailed and costly for practical work. If the persons we use to calibrate items are not too unsymmetrically distributed in ability and not too far off target so that the impact of their ability distribution can be more or less summarized by its mean and variance, then we can use a very simple and workable method for estimating item difficulties. This method, called PROX, was first suggested by Leslie Cohen in 1973 (see Wright and Douglas, 1977a; Wright, 1977).

1.6 A SIMPLE USEFUL ESTIMATION PROCEDURE

Three methods of parameter estimation will be used in this book. The general unconditional method called UCON requires a computer and a computer program such as BICAL (Wright and Mead, 1976). UCON is discussed and illustrated in Chapter 3. A second method called UFORM, which can be done by hand with the help of the simple tables given in Appendix C, is discussed and applied in Chapter 7. The third method, PROX, is completely manageable by hand. In addition the simplicity of PROX helps us to see how the Rasch model works to solve measurement problems. The derivations of the UFORM and PROX equations are given in Wright and Douglas (1975a, 1975b).

PROX assumes that person abilities (β_v) are more or less normally distributed with mean M and standard deviation σ and that item difficulties (δ_i) are also more or less normally distributed with average difficulty H and difficulty standard deviation ω.

If

$$\beta_v \sim N(M, \sigma^2)$$

and

$$\delta_i \sim N(H, \omega^2),$$

then for any person v with person score r_v on a test of L items it follows that

$$b_v = H + X\ell n\,[r_v/(L - r_v)] \qquad [1.6.1]$$

and for any item i with item scores s_i in a sample of N persons it follows that

$$d_i = M + Y\ell n\,[(N - s_i)/s_i] \qquad [1.6.2]$$

The coefficients X and Y are expansion factors which respond in the case of X to the difficulty dispersion of items and in the case of Y to the ability dispersion of persons. In particular

$$X = (1 + \omega^2/2.89)^{1/2} \qquad [1.6.3]$$

and

$$Y = (1 + \sigma^2/2.89)^{1/2} \qquad [1.6.4]$$

The value $2.89 = 1.7^2$ comes from the scaling factor 1.7 which brings the logistic ogive into approximate coincidence with the normal ogive. This is because the logistic ogive for values of 1.7z is never more than one percent different from the normal ogive for values of z.

The estimates b_v and d_i have standard errors

$$SE(b_v) = X\,[L/r_v(L - r_v)]^{1/2} \qquad [1.6.5]$$

$$SE(d_i) = Y\,[N/s_i(N - s_i)]^{1/2} \qquad [1.6.6]$$

This estimation method can be applied directly to observed item scores (s_i) by calculating the sample score logit of item i as

$$x_i = \ell n\ [(N - s_i)/s_i] \qquad [1.6.7]$$

and the item score logit of person v as

$$y_v = \ell n\ [r_v/(L - r_v)] \qquad [1.6.8]$$

The expansion factors X and Y are then estimated by the expressions

$$X = [(1 + U/2.89)/(1 - UV/8.35)]^{1/2}\ , \qquad [1.6.9]$$

for the person logit expansion factor and

$$Y = [(1 + V/2.89)/(1 - UV/8.35)]^{1/2}\ , \qquad [1.6.10]$$

for the item logit expansion factor.

In these expressions $2.89 = 1.7^2$ and $8.35 = 2.89^2 = 1.7^4$ and

$$U = (\sum_{i}^{L} x_i^2 - Lx.^2)/(L - 1), \qquad [1.6.11]$$

the item logit variance and

$$V = (\sum_{v}^{N} y_v^2 - Ny.^2)/(N - 1), \qquad [1.6.12]$$

the person logit variance.

To complete this estimation, we set the test center at zero so that H = 0. Then

$$d_i = M + Yx_i = Y(x_i - x.) \qquad [1.6.13]$$

for each item difficulty, and

$$b_v = H + Xy_v = Xy_v \qquad [1.6.14]$$

for each person ability.

Standard errors are

$$SE(d_i) = Y[N/s_i(N - s_i)]^{1/2} \simeq 2.5/N^{1/2} \qquad [1.6.15]$$

and

$$SE(b_v) = X[L/r_v(L - r_v)]^{1/2} \cong 2.5/L^{1/2} \qquad [1.6.16]$$

Finally the estimated person sample mean and standard deviation become

$$M \approx - Yx\ . \qquad [1.6.17]$$

$$\sigma \approx \quad 1.7(Y^2 - 1)^{1/2} \qquad [1.6.18]$$

Once we have estimated b_v and d_i we can use them to obtain the difference between what the model predicts and the data we have actually observed. These residuals from the model are calculated by estimating the model expectations at each x_{vi} from b_v and d_i and subtracting this expectation from the x_{vi} which was observed. The model expectation for x_{vi} is

$$E\{x_{vi}\} = \pi_{vi}$$

with model variance

$$V\{x_{vi}\} = \pi_{vi}(1 - \pi_{vi})$$

where

$$\pi_{vi} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)] \quad .$$

A standardized residual would be

$$z_{vi} = (X_{vi} - \pi_{vi})/[\pi_{vi}(1 - \pi_{vi})]^{1/2} \quad . \qquad [1.6.19]$$

If the data fit the model this standardized residual ought to be distributed more or less normally with mean zero and variance one.

If we estimate π_{vi} from p_{vi} where

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)] \qquad [1.6.20]$$

then we can use the error distributions

$$z_{vi} \sim N(0,1) \qquad \text{and}$$

$$z_{vi}^2 \sim \chi_1^2$$

as guidelines for evaluating the extent to which any particular set of data can be managed by our measurement model.

We can calculate the sum of their squared residuals z_{vi}^2 for each person. According to the model this sum of squared normal deviates should approximate a chi-square distribution with about $(L - 1)$ degrees of freedom. This gives us a chi-square statistic

$$\sum_{i}^{L} z_{vi}^2 = C_v^2 \sim \chi_{f_v}^2 \qquad [1.6.21]$$

with degrees of freedom

$$f_v = (L - 1)(N - 1)/N \qquad [1.6.22]$$

and a mean square statistic

$$v_v = C_v^2 / f_v \sim F_{f_v, \infty} \qquad [1.6.23]$$

which approximates an F-distribution when the person's responses fit the model.

The sum of squared residuals for each item can be used in the same way to evaluate item fit. For items

$$\sum_{\nu}^{N} z_{\nu\iota}^{2} = C_{\iota}^{2} \sim \chi_{f_{\iota}}^{2} \qquad [1.6.24]$$

with

$$f_{\iota} = (N - 1)(L - 1)/L \qquad [1.6.25]$$

and

$$v_{\iota} = C_{\iota}^{2} / f_{\iota} \sim F_{f_{\iota}, \infty}. \qquad [1.6.26]$$

Finally, since $x_{\nu\iota}$ can only equal one or zero, we can use the definition of $p_{\nu\iota}$ given in Equation 1.6.20 to calculate $z_{\nu\iota}$ and $z_{\nu\iota}^{2}$ directly as

$$z_{\nu\iota} = (2x_{\nu\iota} - 1) \exp [(2x_{\nu\iota} - 1)(d_{\iota} - b_{\nu})/2] \qquad [1.6.27]$$

and

$$z_{\nu\iota}^{2} = \exp[(2x_{\nu\iota} - 1)(d_{\iota} - b_{\nu})] \quad . \qquad [1.6.28]$$

This relation can also be worked backwards. If we already have a $z_{\nu\iota}^{2}$ and wish to calculate the probability of the observed response $x_{\nu\iota}$ to which it refers in order to decide whether or not that response is too improbable to believe, then we can use

$$P\{x_{\nu\iota} \mid b_{\nu}, d_{\iota}\} = 1/(1 + z_{\nu\iota}^{2}). \qquad [1.6.29]$$

In contrast with the $p_{\nu\iota}$ of Equation 1.6.20 which is the estimated probability of a correct answer, the probability of Equation 1.6.29 applies to $x_{\nu\iota}$ whatever value it takes, whether $x_{\nu\iota} = 1$ for a correct answer or $x_{\nu\iota} = 0$ for an incorrect one.

1.7 HOW TRADITIONAL TEST STATISTICS APPEAR IN RASCH MEASUREMENT

Sections 1.1, 1.2 and 1.3 discuss the purpose of tests, the use of test scores and the problems of generality and linearity in making measures. Sections 1.4, 1.5 and 1.6 describe a simple and practical solution to these measurement problems. Because the mathematics are new it might seem that using the Rasch model will take us far away from the traditional item statistics with which we are familiar. This is not so.

Applying the Rasch model in test development gives us new versions of the old statistics. These new statistics contain all of the old familiar information, but in a form which solves most of the measurement problems that have always beset traditional test construction. To show this we will examine the three most common traditional item and person statistics and see how closely they relate to their corresponding Rasch measurement statistics.

The Item P-Value

The most familiar traditional item statistic is the item "p-value." This is the proportion of persons in a specified sample who get that item correct. The PROX estimation equation (1.6.2) gives us a convenient way to formulate the relationship between the traditional item p-value and Rasch item difficulty. If the p-value for item i is expressed as

$$p_i = s_i/N$$

in which s_i is the number of persons in the sample of N persons who answered item i correctly, then the PROX estimated Rasch item difficulty is

$$d_i = M + (1 + \sigma^2/2.89)^{1/2} \ln [(1 - p_i)/p_i] \qquad [1.7.1]$$

Equation 1.7.1 shows that the Rasch item difficulty d_i is in a one-to-one relation with the item p-value represented by p_i. It also shows that this one-to-one relation is curvilinear and involves the ability mean M and variance σ^2 of the calibrating sample.

What the Rasch model does is to use the logit function

$$\ln [(1 - p_i)/p_i]$$

to transform the item p-value which is not linear in the implied variable into a new value which is. This new logit value expresses the item difficulty on an equal interval scale and makes the subsequent correction of the item's p-value for the ability mean M and variance σ^2 of the calibrating sample easy to accomplish.

This correction is made by scaling the logit to remove the effects of sample variance σ^2 and translating this scaled logit to remove the effects of sample mean M. The resulting Rasch item difficulties are not only on an equal interval scale but they are also freed of the observed ability mean and variance of the calibrating sample. Just as the item p-value p_i has a binomial standard error of

$$SE(p_i) = [p_i(1 - p_i)/N]^{1/2} \qquad [1.7.2]$$

so the PROX item difficulty d_i has its own closely related standard error of

$$SE(d_i) = (1 + \sigma^2/2.89)^{1/2} [1/Np_i(1 - p_i)]^{1/2} \quad . \qquad [1.7.3]$$

But there are two important differences between Equations 1.7.2 and 1.7.3. Unlike the p-value standard error in Equation 1.7.2, the Rasch standard error in Equation 1.7.3 is corrected for the ability variance σ^2 of the calibrating sample. The second difference between these two formulations is more subtle, but even more important.

The traditional item p-value standard errors in Equation 1.7.2 are maximum in the middle at p-values near one-half and zero at the extremes at p-values of zero or one. This makes it appear that we know the most about an item, that is have the smallest standard error for its p-value when, in fact, we actually know the least. This is because the item p-value is focused on the calibrating sample as well as on the item. As the sample goes off target for the item, the item p-value nears zero or one and its standard error nears zero. This assures us that the item p-value for this particular sample is extreme but it

tells us nothing else about the item. Thus even though our knowledge of the item's p-value is increasing our information concerning the actual difficulty of the item is decreasing. When item p-values are zero or one, the calibrating sample which was intended to tell us how that item works is shown to be too able or too unable to interact with the item. We know exactly in which direction to look for the item difficulty, but we have no information as to where in that direction it might be.

In contrast, the Rasch standard error for d_i varies in a more reasonable manner. The expression $p_i (1 - p_i)$ which goes to zero as p_i goes to zero or one, appears in the denominator of Equation 1.7.3 instead of in the numerator, as it does in Equation 1.7.2. Therefore, the Rasch standard error is smallest at $p_i = .5$, where the sample is centered on the item and thus gives us the most information about how that item functions. At the extremes, however, where we have the least information, the Rasch standard error goes to infinity reminding us that we have learned almost nothing about that item from this sample.

The Item Point-Biserial

The second most widely used traditional item statistic is the point biserial correlation between the sampled persons' dichotomous responses to an item and their total test scores. The item point-biserial has two characteristics which interfere with its usefulness as an index of how well an item fits with the set of items in which it appears. First, there is no clear basis for determining what magnitude item point-biserial establishes item acceptability. Rejecting the statistical hypothesis that an item point-biserial is zero does not produce a satisfactory statistical criterion for validating an item. The second interfering characteristic is that the magnitude of the point-biserial is substantially influenced by the score distribution of the calibrating sample. A given item's point-biserial is largest when the persons in the sample are spread out in scores and centered on that item. Conversely as the variance in person scores decreases or the sample level moves away from the item level, so that the p-value approaches zero or one, the point-biserial decreases to zero regardless of the quality of the item.

The Rasch statistic that corresponds in meaning to the item point-biserial is the item's mean square residual given in Equation 1.6.26. This mean square residual is not only sensitive to items which fail to correlate with the test score, but also to item point-biserials which are unexpectedly large. This happens, for example, when an additional and unmodelled variable produces a local interaction between a unique feature of the item in question and a corresponding idiosyncrasy among some members of the calibrating sample.

In contrast with the point-biserial, the Rasch item mean square residual has a useful statistical reference distribution. The reference value for testing the statistical hypothesis that an item belongs in the test is a mean square of one with a standard error of $(2/f)^{1/2}$ for f degrees of freedom. Thus the extent to which an observed mean square exceeds the expected value of one can be tested for its statistical significance at whatever significance level is considered useful.

The Rasch item mean square is also very nearly indifferent to the ability distribution of the calibrating sample. This provides a test of item fit which is focused on just those sample and item characteristics which remain when the modelled values for item difficulty and person abilities are removed.

The Person Test Score

The most familiar traditional person statistic is test score, the number of correct answers the person earns on the test taken. Once again we can use the PROX estimation procedure to show the connection between the traditional test score and Rasch person ability. Using the PROX estimation equation (1.6.1) we have

$$b_v = H + (1 + \omega^2/2.89)^{1/2} \ell n\ [r_v/(L - r_v)] \qquad [1.7.4]$$

with a standard error of

$$SE(b_v) = (1 + \omega^2/2.89)^{1/2}\ [L/r_v(1 - r_v)]^{1/2} \qquad [1.7.5]$$

in which

r_v = the test score of person v,
L = the number of items in the test,
H = the average difficulty level of the test and
ω^2 = the variance in difficulties of the test items.

As with the item p-values we see the logit function transforming the person scores which are not linear in the variable they imply into an approximately linear metric. We also see this logit being scaled for test width, which is represented in Equation 1.7.4 by the item difficulty variance ω^2, and then being shifted to adjust for test difficulty level H so that the resulting estimated person ability is freed from the local effects of the test and becomes a test-free measure.

The standard error of this measure is minimum at scores near 50 percent correct, where we have the most information about the person, and goes to infinity at scores of zero and 100 percent, where we have the least information about the person.

While traditional test practices almost always emphasize the analysis of item validity, hardly any attention is ever given to the validity of the pattern of responses leading to a person score. As far as we know no one calculates a person point-biserial coefficient in order to determine the relationship between the responses that person gives to each item and the supposedly relevant item p-values. This would be a reasonable way to apply the traditional point-biserial correlation coefficient to the supervision of person score validity.

The Rasch approach to person score validity is outlined in Equations 1.6.19 through 1.6.23 and discussed and illustrated at length in Chapters 4 and 7.

There are other connections that can be made between traditional test statistics and Rasch statistics. We could review here the various ways that traditional test reliability and validity, norm referencing, criterion referencing, form equating and mastery testing are handled in Rasch measurement. But each of these topics deserves a thorough discussion and that, in fact, is the purpose of the chapters which follow. Our next step now is to see how the PROX estimation procedure works to solve a simple problem in test construction.

2 ITEM CALIBRATION BY HAND

2.1 INTRODUCTION

This chapter describes and illustrates in detail an extremely simple procedure for the Rasch calibration of test items. The procedure, called PROX, approximates the results obtained by more elaborate and hence more accurate procedures extremely well. It achieves the basic aims of Rasch item analysis, namely linearization of the latent scale and adjustment for the local effects of sample ability distribution. The assumption which makes PROX simple is that the effects on item calibration of sample ability distribution can be adequately accounted for by just a mean and standard deviation. This assumption makes PROX so simple that it can easily be applied by hand.

In practice, it will often be convenient to let item calibration be done by computer. However, PROX provides an opportunity to illustrate Rasch item analysis in minute detail, thereby exposing to complete comprehension the process involved, and, where computing facilities are remote or it is urgent to check computer output for plausibility, then PROX provides a method for calibrating items which requires nothing more than the observed distributions of item score and person score, a hand calculator (or adding machine) and paper and pencil.

The data for illustrating PROX come from the administration of the 18-item Knox Cube Test, a subtest of the Arthur Point Scale (Arthur, 1947) to 35 students in Grades 2 to 7. Our analysis of these data shows how Rasch item analysis can be useful for managing not only the construction of national item banks but also the smallest imaginable measurement problem, i.e., one short test given to one roomful of examinees.

Using student correct/incorrect responses to each item of the test, we work out in detail each step of the procedure for PROX item analysis. Then, Chapter 3 reviews comparable computer analyses of the same data by both the PROX procedure and the more accurate UCON procedure used in most computer programs for Rasch item analysis. These detailed steps offer a systematic illustration of the item analysis procedure with which to compare and by which to understand computer outputs. They also demonstrate the ease of hand computations using PROX (PROX is derived and described at length in Wright and Douglas, 1975b, 1976, 1977a; Cohen, 1976, and Wright, 1977). Finally, they illustrate the empirical development of a latent trait or variable. Each step moves from the observed data toward the inferred variable, from the confines of the observed test-bound scores to the reaches of the inferred test-free measurements.

2.2. THE KNOX CUBE TEST

While the Arthur Point Scale covers a variety of mental tasks, the Knox Cube Test implies a single latent trait. Success on this subtest requires the application of visual attention and short-term memory to a simple sequencing task. It appears to be free from school-related tasks and hence to be an indicator of nonverbal intellectual capacity.

The Knox Cube Test uses five one-inch cubes. Four of the cubes are fixed two inches apart on a board, and the fifth cube is used to tap a series on the other four. The four attached cubes will be referred to, from left to right, as "1," "2," "3" and "4" to avoid confusion when specifying any particular series to be tapped. In the original version of the test used for this example, there are 18 such series going from the two-step sequences (1-4) and (2-3) to the seven-step sequence (4-1-3-4-2-1-4). Usually, a subject is administered this test twice with another subtest from the battery intervening. However, we need use only the first administration for our analysis.

The 18 series are given in Figure 2.2.1. These are the 18 "items" of the test. Note that Items 1 and 2 require a two-step sequence; Items 3 through 6, a three-step sequence; Items 7 through 10, a four-step sequence; Items 11 through 13, a five-step sequence, Items 14 through 17, a six-step sequence; and Item 18, a seven-step sequence.

FIGURE 2.2.1

ITEM NAME AND TAPPING ORDER FOR THE KNOX CUBE TEST

ITEM NAME	TAPPING ORDER						
1	1	4					
2	2	3					
3	1	2	4				
4	1	3	4				
5	2	1	4				
6	3	4	1				
7	1	4	3	2			
8	1	4	2	3			
9	1	3	2	4			
10	2	4	3	1			
11	1	3	1	2	4		
12	1	3	2	4	3		
13	1	4	3	2	4		
14	1	4	2	3	4	1	
15	1	3	2	4	1	3	
16	1	4	2	3	1	4	
17	1	4	3	1	2	4	
18	4	1	3	4	2	1	4

2.3 THE DATA FOR ITEM ANALYSIS

The responses of 35 students to a single administration of the 18 item Knox Cube Test are given in Table 2.3.1. These responses are arranged in a person-by-item data matrix. A correct response by a student to an item is recorded as a 1, and an incorrect response as a 0. The items have been listed across the top in the order of administration.

Student scores, the number of correct responses achieved by each student, are given at the end of each row in the last column on the right. Item scores, the total number of correct responses to each item, are given at the bottom of each column.

Inspection of Table 2.3.1 shows that the order of administration is very close to the order of difficulty. Items 1, 2 and 3 are answered correctly by all students. A second, slightly greater, level of difficulty is observed in Items 4 through 9. Then Items 10 and 11 show a sharp increase in difficulty. Items 12 through 17 are answered correctly by only a few students, and no student succeeds on Item 18. Only 12 students score successfully at least once on Items 12 through 17, and only five of these students do one or more of the six-tap items successfully.

2.4 CALIBRATING ITEMS AND MEASURING PERSONS

The general plan for accomplishing item analysis begins with editing the data in Table 2.3.1 to remove persons and items for which no definite estimates of ability or difficulty can be made, i.e., those with all correct or all incorrect responses. This means that Person 35 and Items 1, 2, 3 and 18 must be set aside, leaving 34 persons and 14 items for analysis. Then the remaining information about persons and items is summarized into a distribution of person scores and a distribution of item scores.

Next these score distributions are rendered as proportions of their maximum possible value and their frequency of occurrence is recorded. The proportions are then converted to log odds, or logits, by taking for items the natural log of the proportion incorrect divided by the proportion correct and for persons the natural log of the proportion of successes divided by the proportion of failures. This converts proportions, which are bounded by 0 and 1, to a new scale which extends from $-\infty$ to $+\infty$ and is linear in the underlying variable.

For "item difficulty" this variable increases with the proportion of incorrect responses. For "person ability" it increases with the proportion of correct responses. The mean and variance for each distribution of logits are then computed, and the mean item logit is used to center the item logits at zero. This choice of origin for the new scale is inevitably arbitrary but must be made. Basing it on items rather than on persons and placing it in the center of the current items is natural and convenient.

The item logit and person logit variances are used to calculate two expansion factors, one for items and one for persons. These factors are used to calculate the final sample-free item difficulties and test-free person abilities. They are needed because the apparent relative difficulties of the items depend upon how dispersed in ability the sample of persons is. The more dispersed the persons, the more similar in difficulty will items appear. This is also true for apparent ability. The more dispersed the test in item difficulty, the more similar in ability will persons appear. These effects of sample spread and test width must be removed from the estimates of item difficulty and person ability, if these estimates are to be made sample-free and test-free.

Finally, the standard errors of these estimates are calculated. The standard errors are needed to assess the precision of the estimates. They depend on the same expansion factors plus the extent to which the item difficulty is centered among the person abilities and the person ability is centered among the item difficulties. The more that items or persons are centered on target, the more precise are their estimates and hence the smaller their standard errors.

TABLE 2.3.1

ORIGINAL RESPONSES OF 35 PERSONS TO 18 ITEMS ON THE KNOX CUBE TEST

PERSON NAME	ITEM NAME 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	PERSON SCORE
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	7
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
3	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	10
4	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	6
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
6	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
7	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	14
8	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
9	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
10	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	11
11	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	8
12	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	8
13	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	10
14	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	11
15	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	13
16	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	10
17	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	9
18	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	11
19	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	9
20	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	11
21	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	12
22	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	12
23	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	12
24	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	14
25	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	5
26	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
27	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	7
28	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	10
29	1	1	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0	10
30	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	9
31	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
32	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	11
33	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	6
34	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	12
35	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ITEM SCORE	35	35	35	32	31	30	31	27	30	24	12	6	7	3	1	1	1	0	

Step 1. Organizing the Data Matrix

The data matrix in Table 2.3.1 has been rearranged in Table 2.4.1 so that person scores are ordered from low to high with their respective proportions given in the right-most column, and item scores are ordered from high to low with their proportions given in the bottom row.

Step 2. Editing the Data Matrix

The data matrix of person-by-item responses in Table 2.4.1 has also been edited by removing all items that were answered correctly by everyone or no one, and by removing all persons who had perfect scores or who had not answered any items correctly.

The boundary lines drawn in Table 2.3.1 show the items and persons removed by the editing process. Items 1, 2 and 3 were removed because they were answered correctly by everyone. Removing these three items then brought about the removal of Person 35 because this person had only these three items correct and hence none correct after these items were removed. Item 18 was removed because no person answered this item correctly.

Editing a data matrix may require several such cycles because removing items can necessitate removing persons and vice versa. For example, had there been a person who had succeeded on all but Item 18, then removal of Item 18 would have left this person with a perfect score on the remaining items and so that person would also have had to be removed.

Why were some items and some persons removed? When no one in a sample of persons gets an item correct, that shows that the item is too difficult for this sample of persons. However, no further information is available as to just how much too difficult it actually is. When everyone gets an item correct, that shows that the item is too easy for these persons, but again, no further information is available as to exactly how much too easy the item actually is. To make a definite estimate for a very easy item we must find at least one measurable person who gets it incorrect, and for a very hard item, at least one measurable person who gets it correct. That is, we must "bracket" the item between persons at least one of whom is more and at least one of whom is less able than the item is difficult. Of course, only one person below a very easy item or above a very hard one does not give a very precise estimate of that item's difficulty.

Thus, we have insufficient data in our example to evaluate the extreme Items 1, 2, 3 and 18. We know that Items 1, 2 and 3 appear very easy and that Item 18 appears to be very hard for these persons, but we do not have enough information to specify definite estimates of the difficulties of these four items.

As for extreme persons, do persons with a zero score know nothing? Are scores of 100% indicative of persons who "know it all" or have they only answered easy questions? To make a definite estimate for a person, we must bracket that person between items that are both easier and harder than the person is able.

The boundary scores of zero and 100%, whether for items or for persons, represent incomplete information. They tell us in which direction to look for an estimate of the person's ability or the item's difficulty, but they do not tell us how far to go in that direction. For sufficient information to make a definite estimate of where the person or

TABLE 2.4.1

EDITED AND ORDERED RESPONSES OF 34 PERSONS TO 14 ITEMS

PERSON NAME	ITEM NAME														EDITED PERSON SCORE	PROPORTION OF 14
	4	5	7	6	9	8	10	11	13	12	14	15	16	17		
25	0	1	0	1	0	0	0	0	0	0	0	0	0	0	2	.14
4	1	0	1	0	1	0	0	0	0	0	0	0	0	0	3	.21
33	1	0	1	0	0	0	1	0	0	0	0	0	0	0	3	.21
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4	.29
27	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4	.29
11	0	1	1	1	1	1	0	0	0	0	0	0	0	0	5	.36
12	1	1	1	0	1	0	1	0	0	0	0	0	0	0	5	.36
17	1	0	1	1	1	1	1	0	0	0	0	0	0	0	6	.43
19	1	1	1	1	1	1	0	0	0	0	0	0	0	0	6	.43
30	1	1	1	1	1	1	0	0	0	0	0	0	0	0	6	.43
2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50
3	1	1	1	1	1	1	0	0	0	1	0	0	0	0	7	.50
5	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50
6	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50
8	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50
9	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50
13	1	1	0	0	1	1	1	1	0	1	0	0	0	0	7	.50
16	1	1	1	1	1	1	0	1	0	0	0	0	0	0	7	.50
26	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50
28	1	1	1	1	1	1	0	1	0	0	0	0	0	0	7	.50
29	1	1	0	1	1	0	1	1	0	0	1	0	0	0	7	.50
31	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50
10	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8	.57
18	1	1	1	1	1	1	1	0	1	0	0	0	0	0	8	.57
14	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8	.57
32	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8	.57
20	1	1	1	1	1	1	1	0	1	0	0	0	0	0	8	.57
21	1	1	1	1	1	1	1	1	1	0	0	0	0	0	9	.64
22	1	1	1	1	1	1	1	1	0	1	0	0	0	0	9	.64
23	1	1	1	1	1	1	1	0	1	0	1	0	0	0	9	.64
34	1	1	1	1	1	1	1	0	0	1	1	0	0	0	9	.64
15	1	1	1	1	1	1	1	1	1	1	0	0	0	0	10	.71
7	1	1	1	1	1	1	1	1	1	1	0	1	0	0	11	.79
24	1	1	1	1	1	1	1	1	1	0	0	0	1	1	11	.79
EDITED ITEM SCORE	32	31	31	30	30	27	24	12	7	6	3	1	1	1		
PROPORTION OF 34	.94	.91	.91	.88	.88	.79	.71	.35	.21	.18	.09	.03	.03	.03		

the item is on the latent variable, we must find some items too easy and some items too hard for these persons, and some persons too smart and others too dumb for these items, so that each item and person is bracketed by observations. Then we can make an estimate of where they are on the variable.

Step 3. Obtaining Initial Item Calibrations

From the edited data matrix in Table 2.4.1, we build a grouped distribution of the 10 different item scores and their logits incorrect, and compute the mean and variance of the distribution of these item logits over the test of 14 items. This is done in Table 2.4.2.

EXPLANATION OF TABLE 2.4.2	NOTATION AND FORMULAE
Column 1 of Table 2.4.2 gives the item names collected into each item score group.	
Column 2 gives the item score which characterizes each item score group. Since there are 10 different item scores in this example, G = 10 and the item score group index i goes from 1 to 10.	$s_i \qquad i = 1, G$
Column 3 gives the frequency of items at each score. The sum of these frequencies over the G = 10 item score group comes to the L = 14 items being calibrated.	f_i $L = \sum_{i}^{G} f_i$
Column 4 converts the item scores into proportions correct among the sample of N = 34 persons.	$p_i = s_i/N$
Column 5 is the conversion of proportion correct p_i into the proportion incorrect $1 - p_i$	$1 - p_i = (N - s_i)/N$
Column 6 is the conversion of this proportion into logits incorrect. Each item score group logit is the natural log of its proportion incorrect divided by its proportion correct. This conversion is facilitated by the values of the logits $\ell n[p/(1-p)]$ given in Table 2.4.3.	$x_i = \ell n [(1 - p_i)/p_i]$
Column 7 is the product of item frequency and logit incorrect.	$f_i x_i$
Column 8 is the product of item frequency and logit incorrect squared.	$f_i x_i^2$

TABLE 2.4.2

GROUPED DISTRIBUTION OF THE 10 DIFFERENT ITEM SCORES OF 34 PERSONS

	1	2	3	4	5	6	7	8	9
ITEM SCORE GROUP INDEX	ITEM NAME	ITEM SCORE	ITEM FREQUENCY	PROPORTION CORRECT	PROPORTION INCORRECT	LOGIT INCORRECT	FREQUENCY x LOGIT	FREQUENCY x LOGIT SQUARED	INITIAL ITEM CALIBRATION
i		s_i	f_i	$p_i = s_i/N$	$1 - p_i$	$x_i = \ell n \left[\frac{1-p_i}{p_i}\right]$	$f_i x_i$	$f_i x_i^2$	$d_i^o = x_i - x.$
1	4	32	1	.94	.06	- 2.75*	- 2.75	7.56	- 2.94
2	5, 7	31	2	.91	.09	- 2.31	- 4.62	10.67	- 2.50
3	6, 9	30	2	.88	.12	- 1.99	- 3.98	7.92	- 2.18
4	8	27	1	.79	.21	- 1.32	- 1.32	1.74	- 1.51
5	10	24	1	.71	.29	- 0.90	- 0.90	0.81	- 1.09
6	11	12	1	.35	.65	+ 0.62	+ 0.62	0.38	+ 0.43
7	13	7	1	.21	.79	+ 1.32	+ 1.32	1.74	+ 1.13
8	12	6	1	.18	.82	+ 1.52	+ 1.52	2.31	+ 1.33
9	14	3	1	.09	.91	+ 2.31	+ 2.31	5.34	+ 2.12
10	15, 16, 17	1	3	.03	.97	+ 3.48	+10.44	36.33	+ 3.29

$\sum_i^{10} f_i = 14$ $\quad N = 34$ $\quad \sum_i^{10} f_i x_i = 2.64$ $\quad \sum_i^{10} f_i x_i^2 = 74.81$

$x. = \sum_i^{10} f_i x_i / \sum_i^{10} f_i = 2.64/14 = 0.19$

$U = (\sum_i^{10} f_i x_i^2 - \sum_i^{10} f_i x^2) / (\sum_i^{10} f_i - 1) = (74.81 - 0.51)/13 = 5.72$

Short-cut U'

14/6 = 2.33 $\quad$ Top = 3.29 x 2.33 = 7.67 $\quad$ Bottom = (- 2.94) + (- 2.50 x 1.33) = − 6.27 $\quad$ $U' = \{2\,[7.67 - (-6.27)]\,/13\}^2 = 4.6$

*These values come from Table 2.4.3 where $\ell n[.06/.94] = -2.75$. Were these calculations made with s_i and N as in $\ell n[(N-s_i)/s_i]$ then $\ell n[2/32] = -2.77$. The difference between − 2.75 and − 2.77 is due to the rounding in $s_i/N = 32/34 = 0.941176 \ldots \simeq 0.94$.

TABLE 2.4.3

LOGITS FROM PROPORTIONS

Logit = ℓn [Proportion/(1 − Proportion)]

PROPORTION[1]	LOGIT	PROPORTION	LOGIT	PROPORTION	LOGIT	PROPORTION	LOGIT
.01	− 4.60	.26	− 1.05	51	0.04	.76	1.15
.02	− 3.89	.27	− 0.99	.52	0.08	.77	1.21
.03	− 3.48	.28	− 0.94	.53	0.12	.78	1.27
.04	− 3.18	.29	− 0.90	.54	0.16	.79	1.32
.05	− 2.94	.30	− 0.85	.55	0.20	.80	1.39
.06	− 2.75	.31	− 0.80	.56	0.24	.81	1.45
.07	− 2.59	.32	− 0.75	.57	0.28	.82	1.52
.08	− 2.44	.33	− 0.71	.58	0.32	.83	1.59
.09	− 2.31	.34	− 0.66	.59	0.36	.84	1.66
.10	− 2.20	.35	− 0.62	.60	0.41	.85	1.73
.11	− 2.09	.36	− 0.58	.61	0.45	.86	1.82
.12	− 1.99	.37	− 0.53	.62	0.49	.87	1.90
.13	− 1.90	.38	− 0.49	.63	0.53	.88	1.99
.14	− 1.82	.39	− 0.45	.64	0.58	.89	2.09
.15	− 1.73	.40	− 0.41	.65	0.62	.90	2.20
.16	− 1.66	.41	− 0.36	.66	0.66	.91	2.31
.17	− 1.59	.42	− 0.32	.67	0.71	.92	2.44
.18	− 1.52	.43	− 0.28	.68	0.75	.93	2.59
.19	− 1.45	.44	− 0.24	.69	0.80	.94	2.75
.20	− 1.39	.45	− 0.20	.70	0.85	.95	2.94
.21	− 1.32	.46	− 0.16	.71	0.90	.96	3.18
.22	− 1.27	.47	− 0.12	.72	0.94	.97	3.48
.23	− 1.21	.48	− 0.08	.73	0.99	.98	3.89
.24	− 1.15	.49	− 0.04	.74	1.05	.99	4.60
.25	− 1.10	.50	− 0.00	.75	1.10		

[1] For person scores this "proportion" becomes the number of correct responses r divided by the number of test items L. Thus the person ability logit is $\ell n[(r/L)/(1 - r/L)] = \ell n\,[r/(L - r)]$.
For item scores this "proportion" becomes the number of incorrect responses $(N - s)$ divided by the sample size N. Thus the item difficulty logit is $\ell n\{[(N - s)/N]/[1 - (N - s)/N]\} = \ell n[(N - s)/s]$.

TABLE 2.4.4

GROUPED DISTRIBUTION OF OBSERVED PERSON SCORES ON 14 ITEMS

1	2	3	4	5	6	7
POSSIBLE SCORE	PERSON FREQUENCY	PROPORTION CORRECT	LOGIT CORRECT	FREQUENCY x LOGIT	FREQUENCY x LOGIT SQUARED	INITIAL PERSON MEASURE
r	n_r	$p_r = r/L$	$y_r = \ell n\left[\frac{p_r}{1-p_r}\right]$	$n_r y_r$	$n_r y_r^2$	$b_r^o = y_r$
1	0	.07	- 2.59*	0.00	0.00	- 2.59
2	1	.14	- 1.82	- 1.82	3.31	- 1.82
3	2	.21	- 1.32	- 2.64	3.48	- 1.32
4	2	.29	- 0.90	- 1.80	1.62	- 0.90
5	2	.36	- 0.58	- 1.16	0.67	- 0.58
6	3	.43	- 0.28	- 0.84	0.24	- 0.28
7	12	.50	0.00	0.00	0.00	0.00
8	5	.57	0.28	1.40	0.39	0.28
9	4	.64	0.58	2.32	1.35	0.58
10	1	.71	0.90	0.90	0.81	0.90
11	2	.79	1.32	2.64	3.48	1.32
12	0	.86	1.82	0.00	0.00	1.82
13	0	.93	2.59	0.00	0.00	2.59

$\sum_r^{13} n_r = 34$ $\quad L = 14$ $\quad \sum_r^{13} n_r y_r = -1.00$ $\quad \sum_r^{13} n_r y_r^2 = 15.35$

$y_\cdot = \sum_r^{13} n_r y_r / \sum_r^{13} n_r = -1.00/34 = -0.03$

$V = \left(\sum_r^{13} n_r y_r^2 - \sum_r^{13} n_r y_\cdot^2\right) / \left(\sum_r^{13} n_r - 1\right) = (15.35 - 0.03)/33 = 0.46$

Short-cut V'

34/6 = 5.67 $\quad$ Top = (1.32 x 2) + 0.90 + (2.67 x 0.58) = 5.09 $\quad$ Bottom = - 1.82 + (-1.32 x 2) + (-0.90 x 2) + (-0.58 x 0.67) = - 6.65

$V' = \{2\,[5.09 - (-6.65)]/33\}^2 = 0.5$

*These values come from Table 2.4.3 where $\ell n[.07/.93] = -2.59$. Were r and L used as in $\ell n[r/(L\text{-}r)]$ then $\ell n[1/13] = -2.56$. The difference between -2.59 and -2.56 is due to the rounding in $r/L = 1/14 = 0.071428\ldots \simeq 0.07$.

The mean and variance for the item logits in Column 6 are then computed from the values in Columns 7 and 8 and given beneath these columns.

$$x. = \sum_{i}^{G} f_i x_i / L$$

$$U = (\sum_{i}^{G} f_i x_i^2 - L x.^2) / (L-1)$$

Column 9 gives the values of Column 6 centered by subtracting their mean. These are the initial item calibrations ready to be corrected for the effect of sample spread.

$$d_i^{\circ} = x_i - x.$$

In this example, the mean and variance have been computed from the values in Columns 7 and 8. Hand calibration can be facilitated even further by a short-cut expression for a standard deviation proposed by Mason and Odeh (1968). To do this, sum the item logits in Column 9 (or Column 6) for the top and bottom sixth of the items ordered by difficulty, and take the square of twice the difference of these sums divided by one less than the number of items.

For the data in Table 2.4.2:

a. One-sixth of the items is 14/6 = 7/3, or 2.33 items at each end.

b. The item logits incorrect for the top three items in Column 9 are 3.29, which times 2.33 is 3.29 x 2.33 = 7.67.

c. The item logit incorrect for the bottom item is -2.94 and for the next two items is -2.50. So, taking the lowest item, -2.94, plus 7/3 - 3/3 = 4/3 of the next two items, gives (-2.94) + (-2.50 x 1.33) = − 6.27.

d. The difference between 7.67 and -6.27 is 13.94.

e. Twice this amount divided by the number of items minus one and squared becomes the variance estimate $[2(13.94)/13]^2 = 4.6$.

This short-cut value of 4.6 is somewhat smaller than 5.7 but the number of items is small and the distribution is flatter than the normal distribution assumed by the short-cut.

Completion of the steps in Table 2.4.2 provides initial values for item difficulties in preparation for the adjustment which will compensate for the effect of sample spread.

Step 4. Obtaining Initial Person Measures

In Table 2.4.4, we take identical steps with a grouped distribution of person scores in order to obtain the distribution of person score logits and hence initial values for the abilities that go with each possible score on the test.

EXPLANATION OF TABLE 2.4.4

Column 1 of Table 2.4.4 gives each possible person score from 1 to 13.

NOTATION AND FORMULAE

$$r = 1, L-1$$

EXPLANATION OF TABLE 2.4.4	NOTATION AND FORMULAE
Column 2 gives the frequency of persons observed at each score. The total number of persons N = 34 equals the sum of these frequencies from r = 1 to r = 13.	n_r $N = \sum_{r}^{L-1} n_r$
Column 3 is the proportion of each score on a test of L = 14 items.	$p_r = r/L$
Column 4 is the logit correct for that proportion using Table 2.4.3.	$y_r = \ell n [p_r/(1 - p_r)]$
Column 5 is the product of person frequency and logit correct.	$n_r y_r$
Column 6 is the product of the person frequency and logit correct squared.	$n_r y_r^2$
Column 7 repeats the values of Column 4 because, as far as this test is concerned, the score logits are already centered by the centering of the item logits. These are the initial person measures prior to correction for test width.	$b_r^o = y_r$

The mean and variance for the distribution of score logits over persons are given at the base of Table 2.4.4, as is the short-cut estimate of the variance.

Note that because we are interested not only in the scores observed in this sample but also in the measurements implied by any possible score which might be observed on this test of 14 items, unobserved scores of 1, 12 and 13 have been added to Table 2.4.4, together with the initial measures for these scores. The measurement model specifies what measures are equivalent to these scores even when no persons in the sample actually earn them.

To summarize the procedure thus far (now letting each item define its own item score group for notational simplicity, so that the item index i now runs from 1 to 14 items instead of from 1 to 10 item score groups):

> For a test of L′ items given to N′ persons, we delete all items no one gets correct and no one gets incorrect, and all persons with none correct and none incorrect until no such items or persons remain.
>
> Letting s_i be the number of persons who got item i correct for i = 1 through L, and n_r be the number of persons who got r items correct, for r = 1 through L-1, we find the mean and variance over items of the log odds **incorrect** answers (or item logits incorrect) in the sample to each of the L items and the mean and variance over persons of the log odds **correct** answers (or score logits correct) on the test by each of the N persons.

Thus we obtain, for each item i, its logit incorrect answers among the sample of N persons,

$$x_i = \ell n[(N - s_i)/s_i]$$

and the mean and variance over L items of these item logits.

$$x. = \sum_i^L x_i / L$$

$$U = \sum_i^L (x_i - x.)^2/(L - 1)$$

And we obtain for each score r its logit correct answers on the test of L items,

$$y_r = \ell n[r/(L-r)]$$

and the mean and variance over N persons of their score logits.

$$y. = \sum_r^{L-1} n_r y_r / N$$

$$V = \sum_r^{L-1} n_r (y_r - y.)^2/(N-1)$$

Now we are ready to adjust the initial calibrations and measures in Tables 2.4.2 and 2.4.4 for the local effects of the person ability distribution of the sample and the item difficulty distribution of the test.

Step 5. Calculating the Expansion Factors[1]

We compute expansion factors for the initial estimates of item calibrations and person measures in order to correct the item calibrations for sample spread and the person measures for test width. From Tables 2.4.2 and 2.4.4 we have U = 5.72 and V = 0.46 (or the short-cut values U′ = 4.6 and V′ = 0.5).

a. The person ability expansion factor due to test width is

$$X = \left[\frac{1+U/2.89}{1-UV/8.35}\right]^{1/2}$$

$$= \left[\frac{1+5.72/2.89}{1-(5.72)(0.46)/8.35}\right]^{1/2}$$

$$= \left[\frac{2.98}{0.68}\right]^{1/2} = 2.09$$

or short-cut value

$$X' = \left[\frac{1+U'/2.9}{1-U'V'/8.4}\right]^{1/2}$$

$$= \left[\frac{1+4.6/2.9}{1-(4.6)(0.5)/8.4}\right]^{1/2}$$

$$= \left[\frac{2.59}{0.73}\right]^{1/2} = 1.9$$

[1] For explanation see Chapter One, Section 1.6.

TABLE 2.4.5

FINAL ESTIMATES OF ITEM DIFFICULTIES FROM 34 PERSONS

ITEM SCORE GROUP	1 ITEM NAME	2 INITIAL ITEM CALIBRATION	3 SAMPLE SPREAD EXPANSION FACTOR	4 CORRECTED ITEM CALIBRATION	5 ITEM SCORE	6 CALIBRATION STANDARD ERROR
i		d_i°	Y	$d_i = Yd_i^\circ$	s_i	SE (d_i)
1	4	− 2.94	1.31	− 3.85	32	.95
2	5, 7	− 2.50	1.31	− 3.28	31	.79
3	6, 9	− 2.18	1.31	− 2.86	30	.70
4	8	− 1.51	1.31	− 1.98	27	.56
5	10	− 1.09	1.31	− 1.43	24	.49
6	11	+ 0.43	1.31	0.56	12	.47
7	13	+ 1.13	1.31	1.48	7	.56
8	12	+ 1.33	1.31	1.74	6	.59
9	14	+ 2.12	1.31	2.78	3	.79
10	15, 16, 17	+ 3.29	1.31	4.31	1	1.33

N = 34

$$SE\ (d_i) = Y[N/s_i(N - s_i)]^{1/2}$$

b. The item difficulty expansion factor due to sample spread is

$$Y = \left[\frac{1+V/2.89}{1-UV/8.35}\right]^{1/2}$$

$$= \left[\frac{1+0.46/2.89}{1-(5.72)\,(0.46)/8.35}\right]^{1/2}$$

$$= \left[\frac{1.16}{0.68}\right]^{1/2} = 1.31$$

or short-cut value

$$Y' = \left[\frac{1+V'/2.9}{1-U'V'/8.4}\right]^{1/2}$$

$$= \left[\frac{1+0.5/2.9}{1-(4.6)\,(0.5)/8.4}\right]^{1/2}$$

$$= \left[\frac{1.17}{0.73}\right]^{1/2} = 1.3$$

Step 6. Correcting Item Calibrations for the Effect of Sample Spread

In Table 2.4.5 we obtain the final corrected item calibrations and their standard errors from the sample spread expansion factor Y.

EXPLANATION OF TABLE 2.4.5	NOTATION AND FORMULAE
Column 1 gives the item name.	
Column 2 repeats the initial item calibrations from Column 9 of Table. 2.4.2. (Recall that when items are grouped by item score, then i runs from 1 to G the number of item score groups instead of from 1 to L indexing the individual items.)	$d_i^o = x_i - x.$ $\quad i = 1,G$
Column 3 is the item difficulty expansion factor Y = 1.31 due to sample spread.	Y
Column 4 is the corrected item calibrations obtained by multiplying each initial value in Column 2 by the expansion factor of 1.31.	$d_i = Yd_i^o$ $= Y(x_i - x.)$
Column 5 reminds us of the number of persons who got the items in each item score group correct.	s_i

TABLE 2.4.6

FINAL ESTIMATES OF PERSON MEASURES FOR ALL POSSIBLE SCORES ON THE 14 ITEM TEST

1	2	3	4	5
POSSIBLE TEST SCORE	INITIAL MEASURE	TEST WIDTH EXPANSION FACTOR	CORRECTED MEASURE	MEASURE STANDARD ERROR
r	b_r^o	X	$b_r = Xb_r^o$	SE (b_r)
1	- 2.59	2.09	– 5.41	2.17
2	- 1.82	2.09	– 3.80	1.60
3	- 1.32	2.09	– 2.76	1.36
4	- 0.90	2.09	– 1.88	1.24
5	- 0.58	2.09	– 1.21	1.17
6	- 0.28	2.09	– 0.59	1.13
7	0.00	2.09	0.00	1.12
8	0.28	2.09	0.59	1.13
9	0.58	2.09	1.21	1.17
10	0.90	2.09	1.88	1.24
11	1.32	2.09	2.76	1.36
12	1.82	2.09	3.80	1.60
13	2.59	2.09	5.41	2.17

L = 14

$$SE\ (b_r) = X\,[L/r(L-r)]^{1/2}$$

EXPLANATION OF TABLE 2.4.4	NOTATION AND FORMULAE
Column 6 is the standard error of the corrected item calibrations.	$SE(d_i) = Y[N/s_i(N - s_i)]^{1/2}$

Step 7. Correcting Person Measures for the Effect of Test Width

In Table 2.4.6 we obtain the final corrected person measures and their standard errors from the test width expansion factor X.

EXPLANATION OF TABLE 2.4.6	NOTATION AND FORMULAE
Column 1 gives all possible scores because we want to have measures available for every possible test score from 1 to L – 1, whatever scores were actually observed.	$r = 1, L-1$
Column 2 repeats the initial person measures from Column 7 of Table 2.4.4.	$b_r^o = y_r$
Column 3 is the person ability expansion factor X = 2.09 due to test width.	X
Column 4 is the corrected person measures obtained by multiplying each initial value in Column 2 by the expansion factor of 2.09.	$b_r = Xb_r^o = Xy_r$
Column 5 is the standard error of the corrected person measures.	$SE(b_r) = X[L/r(L-r)]^{1/2}$

2.5 DISCUSSION

The PROX item analysis procedure has been carefully described not only because it accomplishes item calibration and hence person measurement but also because it embodies in a logical and straightforward manner the simplest possible analysis of the interaction between items and persons. The decisive idea on which this analysis is based is that the probability of success is dominated by the person's ability and the item's difficulty. A more able person is supposed always to have a greater chance of success on any item than is a less able person. Any particular person is supposed always to have a better chance of success on an easy item than on a difficulty one. To the extent this is the case the probability of any person's success on any item can be specified as the consequence of the difference between the person's position on a single variable and the item's position on that same variable. That is the Rasch model for item analysis and test construction and, indeed, the fundamental model implicit in the item analysis of all those who work with unweighted scores (Andersen, 1977).

All the information observed about a person's position on the variable, e.g., his ability, is assumed to be expressed in his responses to the set of items he takes as summarized in the unweighted count of the number of items he gets correct. For item difficulty, the information observed is assumed to be completely contained in the unweighted count of persons in the sample who responded correctly to that item.

Of course, this modeling of the interaction between person and item is an idealization and can only approximate whatever actually happens. All ideas, however, are, in the end, only approximations of, or abstractions from, experience. Their value can only be judged in terms of their usefulness, that is, their demonstrable relevance to the situation under study, and their simplicity. This chapter has illustrated the simplicity and potential convenience of Rasch item analysis. Its utility is testified to by hundreds of applications. Our next task is to show how the same data just analyzed by hand come out when the PROX procedure and its more elaborate and accurate parent procedure, UCON, are applied to them by computer.

3 ITEM CALIBRATION BY COMPUTER

3.1 INTRODUCTION

In this chapter we display and describe computer output for Rasch item calibration using the estimation procedures PROX and UCON (Wright and Panchapakesan, 1969; Wright and Douglas, 1975b, 1977a, 1977b). The Knox Cube Test data analyzed by hand in Chapter 2 are used for illustration. The estimation procedures are performed by the computer program BICAL (Wright and Mead, 1976). The accuracy and utility of the hand calibration described in Chapter 2 are evaluated by comparing the "hand" estimates with those produced by a computer analysis of the same data. Knowing the steps by which these procedures can be applied by hand should facilitate understanding and using the computer output.

We will move through the computer output step by step in order to bring out its organization and use. BICAL produces the output given in Tables 3.2.1 to 3.2.9. A comparison of the item and person statistics from PROX by hand with those from PROX by computer is given in Tables 3.3.1 through 3.3.5.

3.2 BICAL OUTPUT FOR A PROX ANALYSIS OF THE KNOX CUBE TEST DATA

The first page of the output, in Table 3.2.1, recaps the control specifications necessary to apply the 1976 version of BICAL to this calibration job (for details consult the manual that goes with your BICAL program). At the top we begin with the job title, the various control parameters, the record (or card) columns read, the test scoring key and a copy of the first person record read. Finally, the output reports that 18 items and 34 persons went into this analysis.

TABLE 3.2.1

PROGRAM CONTROL SPECIFICATIONS

KNOX CUBE TEST

CONTROL PARAMETERS

NITEM	NGROP	MINSC	MAXSC	LREC	KCAB	SCORE
18	10	1	17	21	1	0

COLUMNS SELECTED

```
         1         2         3         4
1********0*********0*********0**********0*****

   11111111111111111

                                KEY

   11111111111111111

                             FIRST SUBJECT
00111111100000000000
```

NUMBER OF ITEMS 18
NUMBER OF SUBJT 34

The control parameters for this job were:

1.	Number of items (NITEM):	18	There were originally 18 items in the KCT.
2.	Smallest subgroup size (NGROP):	10	Subgroups of at least 10 persons are to be formed for analyzing item fit.
3.	Minimum score (MINSC):	1	The minimum score to be used is 1.
4.	Maximum score (MAXSC):	17	The maximum score to be used is 17.
5.	Record length to be read (LREC):	21	The data comes from the first 21 positions of each person record. [The column select card (listed in Table 3.2.1 under "COLUMN SELECTED") specifies the 18 columns that contain these test responses.]
6.	Calibration procedure (KCAB):	1	The calibration procedure to be used is PROX. [The selection code is: 1 = PROX, 2 = UCON]
7.	Scoring option (SCORE):	0	The data are already scored. [The full control code is: 0 = data to be scored dichotomously according to key supplied; 1 = data are successive integers; 2 = score data "correct", if response value equal to or less than key supplied, else "incorrect"; 3 = score data "correct", if response value equal to or greater than key supplied, else "incorrect".]

Table 3.2.2 gives each item's response frequencies for each response value. This table can accommodate up to five response values as specified by the user. An "unknown" value column records the count of all other values encountered. The final column is for the key. The key marks the value specified as correct when the data is still to be scored. As Table 3.2.2 shows the KCT data was entered in scored form. The appropriate key, therefore, is the vector of '1's shown in Tables 3.2.1 and 3.2.2. Each item is identified on the left by its sequence number in the original order of test items as read into BICAL. A four-character item name can also be used to identify test items. For the KCT we have named the items by the number of taps required.

Table 3.2.2 enables us to examine the observed responses for obvious disturbances to our test plan and will often suggest possible explanations for gross misfits. The distribution of responses over multiple-choice distractors, for example, can reveal the undue influence of particular distractors. The effects of insufficient time show up in the piling up of responses in the UNKN column toward the end of the test. The effects of widespread inexperience in test taking show up in the pile-up of UNKN responses in the first one or two items of the test.

We see again in Table 3.2.2 what we already learned from Table 2.3.1, namely that the first three items are answered correctly by all 34 persons, that Item 18 was not answered correctly by anyone and that there is a rapid shift from largely correct responses to largely incorrect responses between Items 9 and 11. Since ITEM NAME gives the number of taps in the series, we see that this shift occurs when the task moves from a series of four taps up to five taps.

TABLE 3.2.2

RESPONSE FREQUENCIES FOR EACH RESPONSE ALTERNATIVE

SEQ NUM	ITEM NAME	ALTERNATIVE RESPONSE FREQUENCIES 0	1				UNKN	KEY
1	2	0	34	0	0	0	0	1
2	2	0	34	0	0	0	0	1
3	3	0	34	0	0	0	0	1
4	3	2	32	0	0	0	0	1
5	3	3	31	0	0	0	0	1
6	3	4	30	0	0	0	0	1
7	4	3	31	0	0	0	0	1
8	4	7	27	0	0	0	0	1
9	4	4	30	0	0	0	0	1
10	4	10	24	0	0	0	0	1
11	5	22	12	0	0	0	0	1
12	5	28	6	0	0	0	0	1
13	5	27	7	0	0	0	0	1
14	6	31	3	0	0	0	0	1
15	6	33	1	0	0	0	0	1
16	6	33	1	0	0	0	0	1
17	6	33	1	0	0	0	0	1
18	6	34	0	0	0	0	0	1

Table 3.2.3 reports the editing process. It summarizes the work of the editing routine which successively removes person records with zero or perfect scores and items correctly answered by all persons or not answered correctly by any persons, until all such persons or items are detected and set aside. The editing process determines the final matrix of item-by-person responses that is analyzed.

Table 3.2.3 shows that initially there were no persons with perfect or zero scores, and that 18 items entered the run, with no person scoring below 1 or above 17, leaving 34 persons for calibration (The 35th person appearing in Table 2.3.1 had already been removed from the data deck by hand.). Items 1, 2 and 3 are then removed by the editing process because they were answered correctly by all subjects and Item 18 is removed because no one answered it correctly. After this editing the calibration sample still consists of 34 subjects, but now only the 14 items which can be calibrated remain, with the same minimum score of 1 and a new maximum score of 13.

Table 3.2.4 shows the distribution of persons over the KCT scores. The histogram is scaled according to the scale factor printed below the graph. The distribution of person scores gives a picture of how this sample responded to these items. It shows how well the items were targeted on the persons and how relevant the persons selected were for this calibration. For the best calibration, persons should be more or less evenly distributed over a range of scores, around and above the center of the test. In our sample we see a symmetrical distribution around a modal score of 7.

TABLE 3.2.3

THE EDITING PROCESS

KNOX CUBE TEST

NUMBER OF ZERO SCORES		0
NUMBER OF PERFECT SCORES		0
NUMBER OF ITEMS SELECTED		18
NUMBER OF ITEMS NAMED		18
SUBJECTS BELOW	1	0
SUBJECTS ABOVE	17	0
SUBJECTS REMAINING		34
TOTAL SUBJECTS		34

REJECTED ITEMS

ITEM NUMBER	ITEM NAME	ANSWERED CORRECTLY	
1	2	34	HIGH SCORE
2	2	34	HIGH SCORE
3	3	34	HIGH SCORE
18	6	0	LOW SCORE

SUBJECTS DELETED	=	0
SUBJECTS REMAINING	=	34
ITEMS DELETED	=	4
ITEMS REMAINING	=	14
MINIMUM SCORE	=	1
MAXIMUM SCORE	=	13

TABLE 3.2.4

SAMPLE PERSON ABILITY DISTRIBUTION

SCORE	COUNT	PROPORTION	DISTRIBUTION OF ABILITY 2 4 6 8 10 *********0********0********0********0********0
1	0	0.0	
2	1	0.03	x
3	2	0.06	xx
4	2	0.06	xx
5	2	0.06	xx
6	3	0.09	xxx
7	12	0.36	xxxxxxxxxxxx
8	5	0.15	xxxxx
9	4	0.12	xxxx
10	1	0.03	x
11	2	0.06	xx
12	0	0.0	
13	0	0.0	
14	0	0.0	

EACH X = 2.94 PERCENT

Table 3.2.5 shows the distribution of item easiness. The scale is given below the graph. Items 4 through 9 are seen to be fairly easy, with Item 8 the most difficult among them. Item 10 is slightly more difficult. Item 11 is much more difficult, followed by Items 12 and 13, more difficult still. Finally, items 15, 16 and 17 are so difficult that only one person answered these items correctly.

TABLE 3.2.5

TEST ITEM EASINESS DISTRIBUTION

ITEM	COUNT	PROPORTION	DISTRIBUTION OF EASINESS 2 4 6 8 10 *********0*********0*********0*********0*********0
4	32	0.94	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
5	31	0.91	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
6	30	0.88	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
7	31	0.91	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
8	27	0.79	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
9	30	0.88	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
10	24	0.71	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
11	12	0.35	xxxxxxxxxxxxxxxx
12	6	0.18	xxxxxxxxx
13	7	0.21	xxxxxxxxxx
14	3	0.09	xxxx
15	1	0.03	x
16	1	0.03	x
17	1	0.03	x

EACH X = 2.00 PERCENT

Table 3.2.6 gives the estimation information. At the top are the PROX difficulty and ability expansion factors. Notice that these values are identical to those we obtained by hand in Chapter 2. Within the table, the first four columns give the item sequence number, item name, item difficulty and standard error.

TABLE 3.2.6

CALIBRATION BY PROX

DIFFICULTY EXPANSION FACTOR 1.31

ABILITY EXPANSION FACTOR 2.10

SEQUENCE NUMBER	ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
4	3	-3.865	0.833
5	3	-3.294	0.691
6	3	-2.876	0.608
7	4	-3.294	0.691
8	4	-2.007	0.485
9	4	-2.876	0.608
10	4	-1.388	0.430
11	5	0.547	0.410
12	5	1.767	0.514
13	5	1.518	0.485
14	6	2.805	0.691
15	6	4.321	1.160
16	6	4.321	1.160
17	6	4.321	1.160

Table 3.2.7 gives the logit ability measure and its standard error for each score on the KCT and the number of persons in the sample obtaining each score. For each raw score we can see the sample frequency at that score and the ability and standard error implied by that score. The sample ability mean and standard deviation are given at the bottom of the table.

TABLE 3.2.7

MEASUREMENT BY PROX

COMPLETE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PERSON ABILITY	STANDARD ERROR
13	0	5.40	1.51
12	0	3.77	1.11
11	2	2.73	0.94
10	1	1.93	0.86
9	4	1.24	0.81
8	5	0.61	0.78
7	12	0.00	0.78
6	3	-0.61	0.78
5	2	-1.24	0.81
4	2	-1.93	0.86
3	2	-2.73	0.94
2	1	-3.77	1.11
1	0	-5.40	1.51

MEAN ABILITY = -0.06
SD OF ABILITY = 1.14

Table 3.2.8 provides item characteristic curves and fit statistics. The tests of fit include a division of the calibration sample into ability subgroups by score level. Three groups have been made out of the KCT sample, the 10 persons with scores from 1 to 6, the 12 persons at score 7 and the 12 persons with scores from 8 to 13. Control over group size and hence over the number of groups used is asserted through the control parameter NGROP. An evaluation of item difficulty invariance over these ability groups is made by comparing for each item its difficulty estimates over the different groups. The tests of fit are thus sample-dependent. However, if the difficulty estimates they use pass these tests, then those estimates are sample-free as far as that sample is concerned. Of course, successful item fit in one sample does not guarantee fit in another. However, as the ability groups within a given sample are arranged by scores, we do obtain information about the stability of item difficulties over various abilities and therefore can see whether our items are displaying sufficient invariance over these particular ability groups to qualify the items for use as instruments of objective measurement.

TABLE 3.2.8

ITEM CHARACTERISTIC CURVES AND ANALYSIS OF FIT BY PROX

		ITEM CHARACTERISTIC CURVE			DEPARTURE FROM EXPECTED ICC			ANALYSIS OF FIT				
SEQ NUM	ITEM NAME	1ST GROUP	2ND GROUP	3RD GROUP	1ST GROUP	2ND GROUP	3RD GROUP	WITHN GROUP	BETWN GROUP	TOTAL	DISC INDX	POINT BISER
4	3	0.80	1.00	1.00	-0.05	0.02	0.01	0.51	0.33	0.49	1.30	0.40
5	3	0.70	1.00	1.00	-0.08	0.04	0.01	0.65	0.58	0.64	1.35	0.42
6	3	0.70	0.92	1.00	-0.02	-0.03	0.02	0.96	0.26	0.90	1.07	0.40
7	4	0.90	0.83	1.00	0.12	-0.13	0.01	1.52	3.89	1.73	0.48	0.23
8	4	0.40	0.92	1.00	-0.16	0.04	0.04	0.43	1.06	0.48	1.49	0.69
9	4	0.60	1.00	1.00	-0.12	0.05	0.02	0.21	0.95	0.27	1.44	0.61
10	4	0.30	0.75	1.00	-0.14	-0.05	0.08	0.80	1.15	0.83	1.40	0.54
11	5	0.0	0.33	0.67	-0.12	-0.03	0.01	0.70	0.82	0.71	1.04	0.54
12	5	0.0	0.17	0.33	-0.04	0.02	-0.06	0.76	0.35	0.72	0.80	0.42
13	5	0.0	0.0	0.58	-0.05	-0.18	0.14	0.26	2.38	0.44	1.49	0.58
14	6	0.0	0.08	0.17	-0.02	0.03	-0.04	0.98	0.23	0.91	0.81	0.20
15	6	0.0	0.0	0.08	0.00	-0.01	0.02	0.17	0.18	0.17	1.33	0.33
16	6	0.0	0.0	0.08	0.00	-0.01	0.02	0.17	0.18	0.17	1.33	0.33
17	6	0.0	0.0	0.08	0.00	-0.01	0.02	0.17	0.18	0.17	1.33	0.33
SCORE RANGE		1–6	7–7	8–13	N = 10	12	12	31	3	34	DEG OF FRDM	
MEAN ABILITY		-1.74	0.00	1.28				0.25	0.82	0.24	STD ERROR	

In the "Item Characteristic Curve" panel of Table 3.2.8 we have the proportion of correct answers given by each ability group to each item. The score range and mean ability for each group are given at the bottom of each column. We expect these ICCs to increase as we move from left to right, from less able to more able score groups, and for the most part we see that in Table 3.2.8 they do. However, Item 7 does show a rather implausible pattern. A greater proportion of persons get it correct in the lowest score group than in the middle one!

In the middle panel of Table 3.2.8 we have the differences in ICC proportions for each ability group between those observed and those predicted by the Rasch measurement model. Here we can see where the largest proportional departures occur and in which direction they go. Again, Item 7 is out of line with the other items, especially for the lowest ability group.

In the "Analysis of Fit" panel of Table 3.2.8 we have a series of fit mean squares. These fit statistics are mean square standardized residuals for item-by-person responses averaged over persons, and partitioned into two components, one between ability groups and the other within ability groups. These mean squares increase in magnitude away from a reference value of 1 as the observed ICC departs from the expected ICC, i.e., when too many high-ability persons fail an easy item or too many low-ability persons succeed on a difficult one. The statistical significance of large values can be judged by comparing the observed mean squares with their expected value of 1 in terms of the expected standard errors given at the bottom of the table.

The "total" mean square evaluates the general agreement between the variable defined by the item and the variable defined by all other items over the whole sample. Only Item 7 is significantly out of line, with an observed mean square of 1.73, more than three times its expected standard error of 0.24 above its expected value of 1.

The "between-group" mean square evaluates the agreement between the observed item characteristic curve and the best fitting Rasch model curve over the ability subgroups. Again, Item 7 is out of line with a mean square of 3.89, more than three times its standard error of 0.82 above 1.

The "within-group" mean square summarizes the degree of misfit remaining within ability groups after the "between-group" misfit has been removed from the "total". Here, Item 7 shows a misfit of 1.52 against an expected value of 1 and a standard error of 0.25.

The discrimination index shown in the next to last column of Table 3.2.8 describes the linear trend of departures from the model across ability groups expressed around a model value of 1. When this index is near 1, then the observed and expected ICCs are close together over the reference points defined by the ability grouping.

When the index is substantially less than 1, then the observed ICC is flatter than expected and the particular item is failing to differentiate among abilities as well as the other items do. This condition, of course, tends to go with a lower point biserial correlation between item response and total test score. However, the discrimination index is less influenced in its magnitude than the point biserial by how central the item is to the sample or how dispersed in ability the sample is.

When the index is substantially greater than 1, then the item gives the appearance of differentiating abilities more distinctly than the average items in the test. The cause of this unusual "discrimination" must then be investigated. It is almost always found to be caused by a local interaction between a secondary characteristic of the item and a

secondary characteristic of the sample, a sample-dependent condition which, upon identification, is generally judged to be too idiosyncratic to be useful in a general measurement system.

The fit statistics in Table 3.2.8 show that Item 7 misfits both "between" and "within" ability groups and that its item characteristic curve is on the flat side. No other item shows a significant misfit. Item 14 does show a low point biserial but its fit statistics are not out of line and, like Items 15, 16, and 17, its low point biserial is due primarily to its difficulty for these persons. Notice how the magnitude of the biserial correlation varies widely with the level of the ICC quite independently of how well the items fit!

This leaves us with the misfit observed for Item 7. What shall we conclude about this misfit? Is it due to a general flaw in Item 7 or is it due to an interaction between Item 7 and a few aberrant person response patterns? It could be that Item 7 functions satisfactorily with most persons and that the misfit observed here can be traced to the irregular responses of just a few persons. Should that be the case we might decide to retain Item 7 in the test and to question, instead, the plausibility of the response patterns of these few unusual persons. We will discuss item and person fit in detail in Chapter 4.

TABLE 3.2.9

ITEM CALIBRATION SUMMARY
BY PROX

SERIAL ORDER					DIFFICULTY ORDER				
SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ
4	3	-3.87	1.30	0.49	4	3	-3.87	1.30	0.49
5	3	-3.29	1.35	0.64	5	3	-3.29	1.35	0.64
6	3	-2.88	1.07	0.90	7	4	-3.29	0.48	1.73
7	4	-3.29	0.48	1.73	6	3	-2.88	1.07	0.90
8	4	-2.01	1.49	0.48	9	4	-2.88	1.44	0.27
9	4	-2.88	1.44	0.27	8	4	-2.01	1.49	0.48
10	4	-1.39	1.40	0.83	10	4	-1.39	1.40	0.83
11	5	0.55	1.04	0.71	11	5	0.55	1.04	0.71
12	5	1.77	0.80	0.72	13	5	1.52	1.49	0.44
13	5	1.52	1.49	0.44	12	5	1.77	0.80	0.72
14	6	2.80	0.81	0.91	14	6	2.80	0.81	0.91
15	6	4.32	1.33	0.17	16	6	4.32	1.33	0.17
16	6	4.32	1.33	0.17	17	6	4.32	1.33	0.17
17	6	4.32	1.33	0.17	15	6	4.32	1.33	0.17
	MEAN	0.00	1.19	0.62					
	S.D.	3.15	0.31	0.42					

FIT ORDER

SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	POINT BISER
16	6	4.32	1.33	0.17	0.33
17	6	4.32	1.33	0.17	0.33
15	6	4.32	1.33	0.17	0.33
9	4	-2.88	1.44	0.27	0.61
13	5	1.52	1.49	0.44	0.58
8	4	-2.01	1.49	0.48	0.69
4	3	-3.87	1.30	0.49	0.40
5	3	-3.29	1.35	0.64	0.42
11	5	0.55	1.04	0.71	0.54
12	5	1.77	0.80	0.72	0.42
10	4	-1.39	1.40	0.83	0.54
6	3	-2.88	1.07	0.90	0.40
14	6	2.80	0.81	0.91	0.20
7	4	-3.29	0.48	1.73	0.23

Table 3.2.9 summarizes the item calibration information in three useful arrangements. We have there for each item its name, difficulty, discrimination index and total fit mean square listed first by serial order, second by difficulty order, and third by fit order. While in the KCT example we have only a few items to deal with, on longer tests the convenient reordering of these items by difficulty and by fit helps us to find misfitting items and to grasp the pattern of misfit, if there is one. In our example we see again that the item with the greatest misfit, Item 7, is identified for us at the bottom of the third panel of Table 3.2.9.

3.3 COMPARING PROX BY HAND WITH PROX BY COMPUTER

Now we can compare the PROX estimation results for item difficulties and person measures obtained by hand with those produced by computer. The data on PROX by hand for item difficulties and person measures comes from Tables 2.4.5 and 2.4.6 in Chapter 2. The data on PROX by computer come from Tables 3.2.6 and 3.2.7. These data have been compiled into Tables 3.3.1 and 3.3.2.

In Table 3.3.1, each item is listed with its calibration by hand and by computer. The standard error for each item as computed by hand and by computer is also given. The results from PROX by hand and PROX by computer are virtually the same.

TABLE 3.3.1

A COMPARISON OF ITEM CALIBRATIONS AND THEIR STANDARD ERRORS FOR PROX BY HAND AND BY COMPUTER

	CALIBRATION			STANDARD ERROR		
Item	Hand	Computer	Difference	Hand	Computer	Difference
4	−3.9	−3.9	0.0	1.0	0.8	0.2
5	−3.3	−3.3	0.0	0.8	0.7	0.1
6	−2.9	−2.9	0.0	0.7	0.6	0.1
7	−3.3	−3.3	0.0	0.8	0.7	0.1
8	−2.0	−2.0	0.0	0.6	0.5	0.1
9	−2.9	−2.9	0.0	0.7	0.6	0.1
10	−1.4	−1.4	0.0	0.5	0.4	0.1
11	+0.6	+0.5	0.1	0.5	0.4	0.1
12	+1.7	+1.8	−0.1	0.6	0.5	0.1
13	+1.5	+1.5	0.0	0.6	0.5	0.1
14	+2.8	+2.8	0.0	0.8	0.7	0.1
15	+4.3	+4.3	0.0	1.3	1.2	0.1
16	+4.3	+4.3	0.0	1.3	1.2	0.1
17	+4.3	+4.3	0.0	1.3	1.2	0.1
MEAN	0.00	0.00		0.82	0.71	
SD	3.13	3.15		0.29	0.29	

The ability measures and their standard errors are given in Table 3.3.2. Again, the differences between the two methods are minimal. Only the standard errors of measurement show a difference of any magnitude. This difference is due to the use of a more accurate but also more laborious formula in PROX by computer. Thus, with the mild exception of the standard errors of measurement, the very simple PROX by hand and PROX by computer produce virtually the same results.

TABLE 3.3.2

A COMPARISON OF PERSON MEASURES AND THEIR STANDARD ERRORS FOR PROX BY HAND AND BY COMPUTER

	MEASURE			STANDARD ERROR		
Score	Hand	Computer	Difference	Hand	Computer	Difference
1	-5.4	-5.4	0.0	2.2	1.5	0.7
2	-3.8	-3.8	0.0	1.6	1.1	0.5
3	-2.8	-2.7	-0.1	1.4	0.9	0.5
4	-1.9	-1.9	0.0	1.2	0.9	0.3
5	-1.2	-1.2	0.0	1.2	0.8	0.4
6	-0.6	-0.6	0.0	1.1	0.8	0.3
7	0.0	0.0	0.0	1.1	0.8	0.3
8	+0.6	+0.6	0.0	1.1	0.8	0.3
9	+1.2	+1.2	0.0	1.2	0.8	0.4
10	+1.9	+1.9	0.0	1.2	0.9	0.3
11	+2.8	+2.7	0.1	1.4	0.9	0.5
12	+3.8	+3.8	0.0	1.6	1.1	0.5
13	+5.4	+5.4	0.0	2.2	1.5	0.7
MEAN	0.00	0.00		1.42	0.98	
SD	3.08	3.06		0.39	0.25	

3.4 ANALYZING KCT WITH THE UCON PROCEDURE

Now that we have seen how PROX by hand compares with PROX by computer, we can turn to a slightly more elaborate and also more accurate procedure which is not suitable for hand work but is convenient and economical to apply by computer. This is the UCON procedure, developed by Wright and Panchapakesan in 1966 (1969) and further reviewed and tested by Wright and Douglas in 1974 (1975b, 1977a) and Wright and Mead in 1975 (1975, 1976). The computer output from UCON is similar in form to that from PROX. The UCON analysis of the KCT data is shown in Tables 3.4.1 through 3.4.4. Only those tables which contain results different from the PROX analysis are presented.

Table 3.4.1 gives the test items with their UCON calibrations and standard errors. UCON uses PROX item difficulties as its point of departure and these are given in the far right column of the table. Table 3.4.2 gives the UCON ability measure associated with each score and the standard error for each measure. The larger standard errors at scores 7 and 8, for abilities between ± 2 logits are caused by the bimodal distribution of item difficulties shown in Table 3.4.1. Six of the 14 items have difficulties below -3.2 logits, while another six have difficulties greater than +1.8 logits. This leaves only two items to function in the 5 logit range between -3.2 and +1.8 and the standard errors of measurement in that region are accordingly higher.

TABLE 3.4.1

CALIBRATION BY UCON

DIFFICULTY EXPANSION FACTOR 1.31

ABILITY EXPANSION FACTOR 2.10

NUMBER OF ITERATIONS = 7

SEQUENCE NUMBER	ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR	LAST DIFF CHANGE	PROX DIFF
4	3	−4.186	0.816	−0.025	−3.865
5	3	−3.648	0.709	−0.023	−3.294
6	3	−3.220	0.647	−0.021	−2.876
7	4	−3.648	0.709	−0.023	−3.294
8	4	−2.241	0.547	−0.015	−2.007
9	4	−3.220	0.647	−0.021	−2.876
10	4	−1.498	0.489	−0.009	−1.388
11	5	0.760	0.456	0.006	0.547
12	5	2.135	0.556	0.015	1.767
13	5	1.861	0.529	0.014	1.518
14	6	3.214	0.705	0.022	2.805
15	6	4.564	1.076	0.027	4.321
16	6	4.564	1.076	0.027	4.321
17	6	4.564	1.076	0.027	4.321

ROOT MEAN SQUARE = 0.022

TABLE 3.4.2

MEASUREMENT BY UCON

COMPLETE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PERSON ABILITY	STANDARD ERROR
13	0	5.09	1.14
12	0	4.11	0.95
11	2	3.31	0.92
10	1	2.53	0.93
9	4	1.71	0.96
8	5	0.81	1.03
7	12	−0.22	1.07
6	3	−1.19	0.97
5	2	−1.96	0.86
4	2	−2.61	0.81
3	2	−3.21	0.81
2	1	−3.86	0.88
1	0	−4.73	1.10

MEAN ABILITY = −0.16

SD OF ABILITY = 1.45

TABLE 3.4.3

ITEM CHARACTERISTIC CURVES AND ANALYSIS OF FIT BY UCON

		ITEM CHARACTERISTIC CURVE			DEPARTURE FROM EXPECTED ICC			ANALYSIS OF FIT				
SEQ NUM	ITEM NAME	1ST GROUP	2ND GROUP	3RD GROUP	1ST GROUP	2ND GROUP	3RD GROUP	WITHN GROUP	BETWN GROUP	TOTAL	DISC INDX	POINT BISER
4	3	0.80	1.00	1.00	−0.04	0.02	0.00	0.39	0.20	0.37	1.13	0.40
5	3	0.70	1.00	1.00	−0.06	0.03	0.01	0.55	0.37	0.54	1.17	0.42
6	3	0.70	0.92	1.00	0.01	−0.04	0.01	0.97	0.26	0.91	0.97	0.40
7	4	0.90	0.83	1.00	0.14	−0.14	0.01	1.72	4.68	1.98	0.60	0.23
8	4	0.40	0.92	1.00	−0.09	0.03	0.03	0.45	0.45	0.45	1.20	0.69
9	4	0.60	1.00	1.00	−0.09	0.05	0.01	0.20	0.60	0.24	1.22	0.61
10	4	0.30	0.75	1.00	−0.04	−0.03	0.05	0.82	0.46	0.79	1.10	0.54
11	5	0.0	0.33	0.67	−0.06	0.06	−0.01	0.80	0.47	0.78	0.96	0.54
12	5	0.0	0.17	0.33	−0.02	0.08	−0.06	0.99	0.76	0.97	0.82	0.42
13	5	0.0	0.0	0.58	−0.02	−0.11	0.13	0.30	1.49	0.41	1.34	0.58
14	6	0.0	0.08	0.17	−0.01	0.05	−0.04	1.39	0.72	1.33	0.82	0.20
15	6	0.0	0.0	0.88	0.00	−0.01	0.01	0.14	0.07	0.13	1.08	0.33
16	6	0.0	0.0	0.08	0.00	−0.01	0.01	0.14	0.07	0.13	1.08	0.33
17	6	0.0	0.0	0.08	0.00	−0.01	0.01	0.14	0.07	0.13	1.08	0.33
SCORE RANGE		1–6	7–7	8–13	N = 10	12	12	31	3	34	DEG OF FRDM	
MEAN ABILITY		−2.30	−0.22	1.67				0.25	0.82	0.24	STD ERROR	

Table 3.4.3 gives the observed Item Characteristic Curve shown in Table 3.2.8, the departures of this ICC from the model ICC as expected by UCON estimates and the fit mean squares resulting from the UCON analysis. Table 3.4.4 summarizes the UCON calibration in the same form as Table 3.2.9 summarizes the PROX calibration.

TABLE 3.4.4

ITEM CALIBRATION SUMMARY BY UCON

SERIAL ORDER					DIFFICULTY ORDER				
SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ
4	3	-4.19	1.13	0.37	4	3	-4.19	1.13	0.37
5	3	-3.65	1.17	0.53	5	3	-3.65	1.17	0.53
6	3	-3.22	0.97	0.90	7	4	-3.65	0.60	1.98
7	4	-3.65	0.60	1.98	6	3	-3.22	0.97	0.90
8	4	-2.24	1.20	0.44	9	4	-3.22	1.22	0.23
9	4	-3.22	1.22	0.23	8	4	-2.24	1.20	0.44
10	4	-1.50	1.10	0.79	10	4	-1.50	1.10	0.79
11	5	0.76	0.96	0.77	11	5	0.76	0.96	0.77
12	5	2.14	0.82	0.97	13	5	1.86	1.34	0.40
13	5	1.86	1.34	0.40	12	5	2.14	0.82	0.97
14	6	3.21	0.82	1.33	14	6	3.21	0.82	1.33
15	6	4.56	1.08	0.13	16	6	4.56	1.08	0.13
16	6	4.56	1.08	0.13	17	6	4.56	1.08	0.13
17	6	4.56	1.08	0.13	15	6	4.56	1.08	0.13
MEAN		0.00	1.04	0.65					
S.D.		3.44	0.19	0.53					

FIT ORDER

SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	POINT BISER
16	6	4.56	1.08	0.13	0.33
17	6	4.56	1.08	0.13	0.33
15	6	4.56	1.08	0.13	0.33
9	4	-3.22	1.22	0.23	0.61
4	3	-4.19	1.13	0.37	0.40
13	5	1.86	1.34	0.40	0.58
8	4	-2.24	1.20	0.44	0.69
5	3	-3.65	1.17	0.53	0.42
11	5	0.76	0.96	0.77	0.54
10	4	-1.50	1.10	0.79	0.54
6	3	3.22	0.97	0.90	0.40
12	5	2.14	0.82	0.97	0.42
14	6	3.21	0.82	1.33	0.20
7	4	3.65	0.60	1.98	0.23

In Chapter 2 we demonstrated the feasibility of hand calibration and showed the computation in detail. This was done to provide a basis for understanding the computer programs which accomplish the same task and their resulting outputs. The comparison of PROX by hand to PROX by computer demonstrates their comparability. In UCON we have a program which provides greater accuracy. Our next step is to compare UCON to PROX.

3.5 COMPARING UCON TO PROX WITH THE KCT DATA

Table 3.5.1 gives the calibrations and standard errors for the KCT data produced by the UCON and PROX methods. The calibration differences between UCON and PROX run about ± .3 logits. The difference between their standard errors is at most ± .1 logits.

TABLE 3.5.1

A COMPARISON OF ITEM CALIBRATIONS AND STANDARD ERRORS FOR UCON AND PROX BY COMPUTER

	CALIBRATION			STANDARD ERROR		
Item	UCON	PROX	Difference	UCON	PROX	Difference
4	-4.2	-3.9	-0.3	0.8	0.8	0.0
5	-3.6	-3.3	-0.3	0.7	0.7	0.0
6	-3.2	-2.9	-0.3	0.6	0.6	0.0
7	-3.6	-3.3	-0.3	0.7	0.7	0.0
8	-2.2	-2.0	-0.2	0.6	0.5	0.1
9	-3.2	-2.9	-0.3	0.6	0.6	0.0
10	-1.5	-1.4	-0.1	0.5	0.4	0.1
11	+0.8	+0.5	0.3	0.5	0.4	0.1
12	+2.1	+1.8	0.3	0.6	0.5	0.1
13	+1.9	+1.5	0.4	0.5	0.5	0.0
14	+3.2	+2.8	0.4	0.7	0.7	0.0
15	+4.6	+4.3	0.3	1.1	1.2	-0.1
16	+4.6	+4.3	0.3	1.1	1.2	-0.1
17	+4.6	+4.3	0.3	1.1	1.2	-0.1
MEAN	0.00	0.00		0.74	0.71	
SD	3.44	3.15		0.22	0.29	

Table 3.5.2 gives the person measures and their standard errors for UCON and PROX. There the differences between UCON and PROX methods run as much as ± .7 logits for the measures.

We see that using the more accurate UCON procedure which takes into account the particular distributions of item difficulties and person abilities does make a tangible difference for the KCT data. As we have seen these KCT items have a distinctly bimodal distribution not well handled by the PROX procedure. Although, these differences between PROX and UCON are never as much as a standard error, and hence could not be

considered statistically significant, nevertheless, they might trouble some practitioners. Their cause, however, is the brevity of this KCT example and the bimodality of its item difficulties. For larger data sets and for more uniform item difficulty distributions, the results of PROX and UCON are virtually indistinguishable.

TABLE 3.5.2

A COMPARISON OF PERSON MEASURES AND STANDARD ERRORS FOR UCON AND PROX BY COMPUTER

	MEASURE			STANDARD ERROR		
Score	UCON	PROX	Difference	UCON	PROX	Difference
1	-4.7	-5.4	0.7	1.1	1.5	-0.4
2	-3.9	-3.8	-0.1	0.9	1.1	-0.2
3	-3.2	-2.7	-0.5	0.8	0.9	-0.1
4	-2.6	-1.9	-0.7	0.8	0.9	-0.1
5	-2.0	-1.2	-0.8	0.9	0.8	0.1
6	-1.2	-0.6	-0.6	1.0	0.8	0.2
7	-0.2	0.0	-0.2	1.0	0.8	0.2
8	0.8	0.6	0.2	1.0	0.8	0.2
9	1.7	1.2	0.5	1.0	0.8	0.2
10	2.5	1.9	0.6	0.9	0.9	0.0
11	3.3	2.7	0.6	0.9	0.9	0.0
12	4.1	3.8	0.3	1.0	1.1	-0.1
13	5.1	5.4	-0.3	1.1	1.5	-0.4
MEAN	0.0	0.0		1.0	1.0	
SD	3.2	3.1		0.1	0.3	

3.6 A COMPUTING ALGORITHM FOR PROX

Here is a concise implementation of the PROX procedure, suitable for computer programming:

1. Edit the binary data matrix of person-by-item responses such that no person has a zero or a perfect score and no item has a zero or a perfect score. This editing may go beyond a single stage when the removal of an item necessitates the removal of some persons, and vice versa. The final outcome is a vector of item scores (s_i) where i goes from 1 to L and a vector of person score frequencies (n_r) where r goes from 1 to L-1.

2. Let: $x_i = \ell n\ [(N - s_i)/s_i]$ [3.6.1]

$$x_. = \sum_i^L x_i/L \qquad [3.6.2]$$

$$y_r = \ell n\ [r/(L - r)] \qquad [3.6.3]$$

$$y_. = \sum_{r}^{L-1} n_r y_r / N \qquad [3.6.4]$$

$$D = \sum_{i}^{L} (x_i - x_.)^2 / 2.89\,(L-1) \qquad [3.6.5]$$

$$B = \sum_{r}^{L-1} n_r (y_r - y_.)^2 / 2.89\,(N-1) \qquad [3.6.6]$$

$$G = BD \qquad [3.6.7]$$

3. Calculate the expansion factors:

$$X = [(1 + D) / (1 - G)]^{1/2} \qquad [3.6.8]$$

$$Y = [(1 + B) / (1 - G)]^{1/2} . \qquad [3.6.9]$$

4. Estimate the item difficulties as:

$$d_i = Y (x_i - x_.), \qquad \text{for } i = 1, L \qquad [3.6.10]$$

5. With standard errors of:

$$SE(d_i) = Y [N/s_i (N - s_i)]^{1/2}. \qquad [3.6.11]$$

6. The ability estimates for this set of items are given by:

$$b_r = X y_r, \qquad \text{for } r = 1, L-1 \qquad [3.6.12]$$

7. With standard errors of:

$$SE(b_r) = X [L/r(L-r)]^{1/2} . \qquad [3.6.13]$$

3.7 THE UNCONDITIONAL PROCEDURE UCON

The Rasch model for binary observations defines the probability of a response $x_{\nu i}$ to item i by person ν as

$$P\{x_{\nu i} \mid \beta_\nu, \delta_i\} = \exp [x_{\nu i}(\beta_\nu - \delta_i)] / [1 + \exp (\beta_\nu - \delta_i)] \qquad [3.7.1]$$

where $x_{\nu i} = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise,} \end{cases}$

β_ν = ability parameter of person ν,

δ_i = difficulty parameter of item i.

The likelihood Λ of the data matrix $((x_{\nu i}))$ is the continued product of Equation [3.7.1] over all values of ν and i, where L is the number of items and N is the number of persons with test scores between 0 and L, since scores of 0 and L lead to infinite ability estimates.

$$\Lambda = \exp [\sum_{\nu}^{N} \sum_{i}^{L} x_{\nu i} (\beta_\nu - \delta_i)] / \prod_{\nu}^{N} \prod_{i}^{L} [1 + \exp (\beta_\nu - \delta_i)] \qquad [3.7.2]$$

Upon taking logarithms and letting

$$\sum_{i}^{L} x_{\nu i} = r_\nu \quad \text{be the score of person } \nu$$

and $$\sum_{\nu}^{N} x_{\nu i} = s_i \quad \text{be the score of item i,}$$

the log likelihood λ becomes

$$\lambda = \ell n\, \Lambda = \sum_{\nu}^{N} r_\nu \beta_\nu - \sum_{i}^{L} s_i \delta_i - \sum_{\nu}^{N} \sum_{i}^{L} \ell n\, [1 + \exp(\beta_\nu - \delta_i)] \,. \qquad [3.7.3]$$

The reduction of the data matrix $((x_{\nu i}))$ to its margins (r_ν) and (s_i) and the separation of $r_\nu \beta_\nu$ and $s_i \delta_i$ in Equation 3.7.3 establish the sufficiency of r_ν for estimating β_ν and of s_i for estimating δ_i, as well as the objectivity of these estimates.

It is important to recognize, of course, that although r_ν and s_i lead to sufficient estimates of β_ν and δ_i they themselves are not satisfactory as estimates. Person score r_ν is not free from the particular item difficulties encountered in the test. Nor is item score s_i free from the ability distribution of the persons who happen to be taking the item. To achieve independence from these local factors requires adjusting the observed r_ν and s_i for the related item difficulty and person ability distributions to produce the test-free person measures and sample-free item calibrations desired.

With the side condition $\sum_{i}^{L} \delta_i = 0$ to restrain the indeterminacy of origin in the response parameters, the first and second partial derivatives of λ with respect to β_ν and δ_i become

$$\frac{\partial \lambda}{\partial \beta_\nu} = r_\nu - \sum_{i}^{L} \pi_{\nu i} \qquad \nu = 1,N \qquad [3.7.4]$$

$$\frac{\partial^2 \lambda}{\partial \beta_\nu^2} = - \sum_{i}^{L} \pi_{\nu i}(1 - \pi_{\nu i}) \qquad [3.7.5]$$

and $$\frac{\partial \lambda}{\partial \delta_i} = -s_i + \sum_{\nu}^{N} \pi_{\nu i} \qquad i = 1,L \qquad [3.7.6]$$

$$\frac{\partial^2 \lambda}{\partial \delta_i^2} = - \sum_{\nu}^{N} \pi_{\nu i}\,(1 - \pi_{\nu i}) \qquad [3.7.7]$$

where $\pi_{\nu i} = \exp(\beta_\nu - \delta_i)/[1 + \exp(\beta_\nu - \delta_i)]$

These are the equations necessary for unconditional maximum likelihood estimation. The solutions for item difficulty estimates in Equations 3.7.6 and 3.7.7 depend on the presence of values for the person ability estimates. Because unweighted test scores are the sufficient statistics for estimating abilities, all persons with identical scores obtain identical ability estimates. Hence, we may group persons by their score, letting

b_r be the ability estimate for any person with score r,
d_i be the difficulty estimate of item i,
n_r be the number of persons with score r,

and write the estimated probability that a person with a score r will succeed on item i as

$$p_{ri} = \exp(b_r - d_i)/[1 + \exp(b_r - d_i)] \ . \qquad [3.7.8]$$

Then $\sum_{\nu}^{N} \pi_{\nu i} \approx \sum_{r}^{L-1} n_r p_{ri}$, as far as estimates are concerned.

A convenient algorithm for computing estimates (d_i) is:

1. Define an initial set of (b_r) as

$$b_r^{(0)} = \ln\left(\frac{r}{L-r}\right) \qquad r = 1, L-1 \qquad [3.7.9]$$

2. Define an initial set of (d_i), centered at $d_\cdot = 0$, as

$$d_i^{(0)} = \ln\left(\frac{N - s_i}{s_i}\right) - \left[\sum_{1}^{L} \ln\left(\frac{N - s_i}{s_i}\right)\right]/L \qquad i = 1, L \qquad [3.7.10]$$

where $b_r^{(0)}$ is the maximum likelihood estimate of β_ν for a test of L equivalent items centered at zero and $d_i^{(0)}$ is the similarly centered maximum likelihood estimate of δ_i for a sample of N equal-ability persons.

3. Apply Newton's method to Equation 3.7.6 to improve each d_i according to ,

$$d_i^{(j+1)} = d_i^{(j)} - \frac{-s_i + \sum_{r}^{L-1} n_r p_{ri}^{(j)}}{-\sum_{r}^{L-1} n_r p_{ri}^{(j)} (1 - p_{ri}^{(j)})} \qquad i = 1, L \qquad [3.7.11]$$

until convergence at $|d_i^{(j+1)} - d_i^{(j)}| < .01$

where $p_{ri}^{(j)} = \exp(b_r - d_i^{(j)})/[1 + \exp(b_r - d_i^{(j)})]$ [3.7.12]

and the current set of (b_r) are given by the previous cycle.

4. Recenter the set of (d_i) at $d_\cdot = 0$.

5. Using this improved set of (d_i), apply Newton's method to Equation 3.7.4 to improve each b_r according to

$$b_r^{(m+1)} = b_r^{(m)} - \frac{r - \sum_{i}^{L} p_{ri}^{(m)}}{-\sum_{i}^{L} p_{ri}^{(m)} (1 + p_{ri}^{(m)})} \qquad r = 1, L-1 \qquad [3.7.13]$$

until convergence at $|b_r^{(m+1)} - b_r^{(m)}| < .01$

where $$p_{ri}^{(m)} = \exp(b_r^{(m)} - d_i)/[1 + \exp(b_r^{(m)} - d_i)] \qquad [3.7.14]$$

and the current set of (d_i) are given by the previous cycle.

6. Repeat steps (3) through (5) until successive estimates of the whole set of (d_i) become stable at

$$\sum_{i}^{L} (d_i^{(k+1)} - d_i^{(k)})^2/L < .0001 , \qquad [3.7.15]$$

which usually takes three or four cycles.

7. Use the reciprocals of the negative square roots defined in Equation 3.7.7 as asymptotic estimates of the standard errors of difficulty estimates,

$$SE(d_i) = [\sum_{r}^{L-1} n_r p_{ri}(1 - p_{ri})]^{-1/2} . \qquad i = 1, L \qquad [3.7.16]$$

Andersen (1973) has shown that the presence of the ability parameters (β_v) in the likelihood equation of this unconditional approach leads to biased estimates of item difficulties (δ_i). Simulations undertaken to test UCON in 1966 indicated that multiplying the centered item difficulty estimates by the coefficient [(L - 1)/L] compensates for most of this bias. (For a discussion and evaluation of the unbiasing coefficient [(L - 1)/L] see Wright and Douglas, 1975b or 1977a).

4 THE ANALYSIS OF FIT

4.1 INTRODUCTION

Procedures for item calibration by hand were given in Chapter 2, and calibration output from computer programs was discussed in Chapter 3. However, these calibration procedures are only part of a complete analysis of a sample of data. The Rasch model makes certain plausible assumptions about what happens when a person takes an item, and a complete analysis must include an evaluation of how well the data fit these assumptions. When, for example, a person answers all the hard items of a test correctly but then misses several easy items, we are surprised by the resulting implausible pattern of responses. While we could examine individual records by eye for their implausibility, in practice we want to put such evaluations on a systematic and manageable basis. We want to be able to be specific and objective in our reactions to implausible observations.

Even if the measurement model tends to fit a particular application, we cannot predict in advance how well new items (or even old ones) will continue to work in every situation in which they might be applied, nor can we know in advance how all persons will always respond. Therefore, if we are serious in our attempts to measure, we must examine every application to see how well each set of responses corresponds to our model expectations. We must evaluate not only the plausibility of the sample of persons' responses, but also the plausibility of each person's responses to the set of items in his test. To do this we must examine the response of each person to each item to determine whether it is consistent with the general pattern of responses observed.

4.2. THE KCT RESPONSE MATRIX

We begin the study of fit analysis by returning to the item-by-person data matrix of the KCT given in Table 2.4.1. In this table we have the edited and ordered responses of 34 persons to 14 KCT items. The editing process removed items answered correctly by everyone or no one, and persons answering correctly all or none of the items. The remaining persons and items have been arranged in order of increasing item and person score.

This item-by-person matrix of 1's and 0's is the complete record of usable person responses to the items of the test. By inspection we see that the increasing difficulty of the KCT items has divided the matrix roughly into two triangles: a lower left triangle dominated by correct responses signified by 1's and an upper right triangle dominated by incorrect responses signified by 0's.

This is the pattern we expect. As items get harder, going from left to right in Table 2.4.1, any particular person's string of successes should gradually peter out and end in a string of failures on the items much too hard for that person. Similarly, when we examine the pattern of responses for any item by proceeding from the bottom of Table 2.4.1 up

that item's column over persons of decreasing ability, we expect the string of successes at the bottom to peter out into failures as the persons become too low in ability to succeed on this item.

From our calibration of the KCT items we have estimates of the item difficulties (d_i) and of the abilities (b_r) which go with the possible scores (r) on this test. In Table 4.2.1 we show the matrix of responses from Table 2.4.1 to which we have added, from our calibration, the item difficulties (d_i) across the bottom and the abilities (b_r) associated with each score down the right column. The item difficulties and score abilities in Table 4.2.1 are those estimated with PROX by hand from Chapter 2.

Notice how Table 4.2.1 is arranged into six sections in order to bring out the pattern of responses. The 14 items are partitioned into the 7 easier and the 7 harder. The 34 persons are partitioned into the 10 scoring **below** seven, the 12 scoring exactly seven and the 12 scoring **above** seven. In the lower left section there are only 1's. Every higher ability person got every easier item correct. In the upper right section there are only 0's. Every lower ability person got every harder item incorrect. But in the other four sections there is a pattern of 1's and 0's that must be analyzed.

When we examine the pattern of responses in these data for unexpected "corrects" and "incorrects," we find that Table 4.2.1 shows several exceptions to a pattern of all 1's followed by all 0's. Of course, we do not expect every single person to fail for the first time at a particular point and then always to continue to do so on all harder items. We expect to find a run of successes and failures leading finally to a run of failures as the items finally become too difficult. However, some of the exceptions in Table 4.2.1 seem to exceed even this expectation. To facilitate their examination we have circled those responses which seem most unexpected given the overall pattern.

The locations of these apparently surprising responses lead us to examine more closely some of the person records in Tables 4.2.1.

For Person 2, the pattern of responses is almost too reasonable: all 1's followed by all 0's.

For Person 29, in contrast, the pattern is quite puzzling; it shows both failures on easy items and success on a hard one.

The expected pattern is the one we see in the records of Persons 12 or 23. Here each record shows a string of 1's with a few adjacent and alternating 1's and 0's, followed by a string of 0's.

Turning to the four most questionable records, we see:

Person 11, failed Item 4 but passed Items 5 through 9 before failing all the remaining items.

Person 17, passed Item 4, missed Item 5, passed Items 6 through 10 and then failed the remaining items.

Person 13, passed Items 4 and 5, missed Items 6 and 7, passed Items 8 through 12 and then missed the remaining ones.

TABLE 4.2.1

RESPONSE PATTERNS OF 34 PERSONS TO 14 ITEMS FROM TABLE 2.4.1

	PERSON NAME	EASY Items →							HARD Items →							SCORE r	PROPORTION OF 14	SCORE ABILITY b_r
		(4)	(5)	(7)	(6)	9	(8)	10	11	13	(12)	(14)	15	16	17			
Low Persons	25	0	1	0	1	0	0	0	0	0	0	0	0	0	0	2	.14	-3.8
	4	1	0	1	0	1	0	0	0	0	0	0	0	0	0	3	.21	-2.8
	33	1	0	1	0	0	0	1	0	0	0	0	0	0	0	3	.21	
	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4	.29	-1.9
	27	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4	.29	
	(11)	0	1	1	1	1	1	0	0	0	0	0	0	0	0	5	.36	-1.2
	(32)	1	1	1	0	1	0	1	0	0	0	0	0	0	0	5	.36	
	(17)	1	0	1	1	1	1	1	0	0	0	0	0	0	0	6	.43	
	19	1	1	1	1	1	1	0	0	0	0	0	0	0	0	6	.43	-0.6
	30	1	1	1	1	1	1	0	0	0	0	0	0	0	0	6	.43	
	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50	
	(3)	1	1	1	1	1	1	0	0	0	(1)	0	0	0	0	7	.50	
	5	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50	
	6	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50	
	8	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50	
	9	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50	
	(13)	1	1	0	0	1	1	1	1	0	(1)	0	0	0	0	7	.50	0.0
	16	1	1	1	1	1	1	0	1	0	0	0	0	0	0	7	.50	
	26	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50	
	28	1	1	1	1	1	1	0	1	0	0	0	0	0	0	7	.50	
	(29)	1	1	0	1	1	0	1	1	0	0	(1)	0	0	0	7	.50	
	31	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7	.50	
High Persons	10	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8	.57	
	18	1	1	1	1	1	1	1	0	1	0	0	0	0	0	8	.57	
	14	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8	.57	0.6
	32	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8	.57	
	20	1	1	1	1	1	1	1	0	1	0	0	0	0	0	8	.57	
	21	1	1	1	1	1	1	1	1	1	0	0	0	0	0	9	.64	
	22	1	1	1	1	1	1	1	1	0	1	0	0	0	0	9	.64	1.2
	23	1	1	1	1	1	1	1	0	1	0	1	0	0	0	9	.64	
	34	1	1	1	1	1	1	1	0	0	1	1	0	0	0	9	.64	
	15	1	1	1	1	1	1	1	1	1	1	0	0	0	0	10	.71	1.9
	7	1	1	1	1	1	1	1	1	1	1	0	1	0	0	11	.79	
	24	1	1	1	1	1	1	1	1	1	0	0	0	1	1	11	.79	2.8
	PROPORTION OF 34	94	91	91	88	88	79	71	35	21	19	09	03	03	03			
	Item Difficulty d_i	-3.9	-3.3	-3.3	-2.9	-2.9	-2.0	-1.4	0.6	1.5	1.7	2.8	4.3	4.3	4.3			

Item difficulties (d_i) and score abilities (b_r) from Tables 2.4.5 and 2.4.6

Person 29, passed Items 4 through 6, missed Items 7 and 8, passed Items 9 through 11, missed Items 12 and 13 and then passed Item 14 before missing all the remaining items.

There are a few other records that might also be examined such as Persons 3 and 12, but as we "eyeball" this small matrix, we can see that the other records are less exceptional.

Now that we have found some instances of possibly irregular responses, we want a systematic way to judge the degree of unexpectedness seen in these response patterns.

The Rasch model bases calibration and measurement on two expectations: (1) that a more able person should always have a greater probability of success on any item than a less able person, and (2) that any person should always be more likely to do better on an easier item than on a harder one. When an observed pattern of responses shows significant deviations from these expectations, we can use the particulars of the model and the person and item estimates to calculate an index of how unexpected any particular person or item record is.

4.3 THE ANALYSIS OF FIT BY HAND

The first step in our analysis of response plausibility or fit is to observe the difference $(b_\nu - d_i)$ between the estimates of ability b_ν and difficulty d_i for each person and item. When this difference is positive, it means that the item should be easy for the person. The more positive the difference, the easier the item and hence the greater our expectation that the person will succeed. Similarly, as the difference between person ability and item difficulty becomes more and more negative, the item should be more and more difficult for that person, and our expectation of his failure increases.

In order to focus our application of these ideas, we have taken from Table 4.2.1 the responses of the six persons with the most implausible patterns to the seven items on which their implausible responses occur. These selected responses comprise Table 4.3.1. With this table we can more easily study the outstanding unexpected "correct" or "incorrect" responses.

To begin with, we can tabulate the number of unexpected responses for each person and item in Table 4.3.1 to arrive at a simple count with which to describe what is occurring. We see that Persons 13 and 29 make the worst showing with three unexpected responses each. However, this simple count does not tell us how to weigh and hence how to judge the degree of unexpectedness in these responses.

One way statisticians think about the outcomes of probabilistic events like dice-rolling, coin-tossing and getting an item correct on a test is to define the expected value of the variable realized in any response $x_{\nu i}$, say of person ν to item i, as the probability $\pi_{\nu i}$ of that response occurring. This is useful because, if we were to obtain response $x_{\nu i}$ a great many times and its genesis were more or less governed by the probability $\pi_{\nu i}$, then we would expect success to occur about $\pi_{\nu i}$ of the time, just as we expect "6" to come up about one-sixth of the time when we roll dice and "heads" to come up about one-half of the time when we toss coins.

TABLE 4.3.1

SELECTED PERSON-TO-ITEM RESPONSES ($x_{\nu i}$) WITH UNEXPECTED RESPONSES CIRCLED

	ITEM							NUMBER OF UNEXPECTED	PERSON
PERSON	4	5	7	6	8	12	14	RESPONSES	ABILITY*
11	(0)	1	1	1	1	0	0	1	−1.2
12	1	1	1	(0)	(0)	0	0	2	−1.2
17	1	(0)	1	1	1	0	0	1	−0.6
3	1	1	1	1	1	(1)	0	1	0.0
13	1	1	(0)	(0)	1	(1)	0	3	0.0
29	1	1	(0)	1	(0)	0	(1)	3	0.0
Number of Unexpected Responses	1	1	2	2	2	2	1	11	
Item Difficulty*	−3.9	−3.3	−3.3	−2.9	−2.0	1.7	2.8		
	"1" expected "0" unexpected					"0" expected "1" unexpected			

*From Tables 2.4.5 and 2.4.6

Our model estimates the probability $\pi_{\nu i}$ of instances of response $x_{\nu i}$ as

$$p_{\nu i} = \exp(b_\nu - d_i)/[1 + \exp(b_\nu - d_i)]$$

where b_ν = the estimated ability measure of person ν

and d_i = the estimated difficulty calibration of item i.

Thus we can use $p_{\nu i}$ as an estimate of the expected value of instances of $x_{\nu i}$.

The same theory tells us that the expected variance of instances of $x_{\nu i}$ is $\pi_{\nu i}(1 - \pi_{\nu i})$ which we can estimate with $p_{\nu i}(1 - p_{\nu i})$. The result is an estimated standard residual $z_{\nu i}$ from any $x_{\nu i}$ of

$$z_{\nu i} = (x_{\nu i} - p_{\nu i})/[p_{\nu i}(1 - p_{\nu i})]^{1/2} \quad . \qquad [4.3.1]$$

To estimate this standard residual $z_{\nu i}$, we subtract from the observed $x_{\nu i}$ its estimated expected value $p_{\nu i}$ and standardize this residual difference by the divisor

$$[p_{\nu i}(1 - p_{\nu i})]^{1/2}$$

which is the estimated binomial standard deviation of such observations. To the extent that our data approximate the model, we expect this estimated residual z_{vi} to be distributed more or less normally with a mean of about 0 and a variance of about 1.

Thus, as a rough but useful criterion for the fit of the data to the model, we can examine the extent to which these standard residuals approximate a normal distribution, i.e.

$$z_{vi} \sim N(0, 1)$$

or their squares approximate a one degree of freedom chi-square distribution, i.e.

$$z_{vi}^2 \sim \chi_1^2 \quad .$$

The reference values of 0 for the mean and 1 for the standard deviation and the reference distributions of N(0, 1) and χ_1^2 help us to see if the estimated standard residuals deviate significantly from their model expectations. This examination of residuals will suggest whether we can proceed to use these items to make measurements, or whether we must do further work on the items and the testing situation to bring them into line with reasonable expectations. It will also indicate when particular persons have failed to respond to the test in a plausible manner.

When a particular squared residual z_{vi}^2 becomes very large, we wonder if something unexpected happened when person v took item i. Of course, a single unexpected response is less indicative of trouble than a string of unexpectedly large values of z_{vi}^2. Then the accumulated impact of these values taken over items for a person or over persons for an item is bound to produce concern for the plausibility of the person's measure or of the item's calibration and hence to put into doubt the meaning of that person's measurement or of that item's calibration.

Since x_{vi} takes only the two values of "0" and "1" we can express these standard residuals in terms of the estimates b_v and d_i.

From Equation 4.3.1 we have

$$z_x = (x - p)/[p(1 - p)]^{1/2} \quad .$$

So when $x = 0$ then $z_o = (-p)/[p(1 - p)]^{1/2} = -[p/(1 - p)]^{1/2}$

and $x = 1$ then $z_1 = (1 - p)/[p(1 - p)]^{1/2} = +[(1 - p)/p]^{1/2}$.

Now since $p = \exp(b - d)/[1 + \exp(b - d)]$

then $p/(1 - p) = \exp(b - d)$

and $(1 - p)/p = \exp(d - b)$.

So $z_o = -\exp[(b - d)/2]$ $\qquad z_o^2 = \exp(b - d)$

and $z_1 = +\exp[(d - b)/2]$ $\qquad z_1^2 = \exp(d - b)$.

or in general

$$z = (2x\text{-}1)\exp[(2x\text{-}1)(d\text{-}b)/2] \qquad [4.3.2]$$

$$z^2 = \exp[(2x\text{-}1)(d\text{-}b)] \qquad [4.3.3]$$

Thus, exp (b - d) indicates the unexpectedness of an incorrect response to a relatively easy item, while exp (d - b) indicates the unexpectedness of a correct response to a relatively hard item. The values of $z_0^2 = \exp(b - d)$ and $z_1^2 = \exp(d - b)$ can be ascertained for each x_{vi} of 0 or 1 and then accumulated over items to evaluate the plausibility of any person measure, or over persons to evaluate the plausibility of any item calibration.

To evaluate the unexpected responses in Table 4.3.1 we replace each instance of an unexpected response by the difference between the ability measure for that person and the difficulty calibration for that item. For Person 11 on Item 4 the unexpected incorrect response associated with a person ability b_v of -1.2 and an item difficulty d_i of -3.9 leads to a difference $(b_v - d_i)$ of $(-1.2) - (-3.9) = +2.7$.

This difference of 2.7 for Person 11 on Item 4 is placed at the location of that unexpected response in the matrix in Table 4.3.2 where we have also computed the differences for each instance of an unexpected response given in Table 4.3.1.

TABLE 4.3.2

ABILITY - DIFFICULTY DIFFERENCES ($b_v - d_i$) FOR UNEXPECTED RESPONSES

	ITEM							
PERSON	4	5	7	6	8	12	14	PERSON ABILITY
11	2.7							-1.2
12				1.7	0.8			-1.2
17		2.7						-0.6
3						1.7		0.0
13			3.3	2.9		1.7		0.0
29			3.3		2.0		2.8	0.0
Item Difficulty	-3.9	-3.3	-3.3	-2.9	-2.0	1.7	2.8	

Since "1" expected "0" unexpected entry is (b - d)

Since "0" expected "1" unexpected entry is (d - b)

Unexpected incorrect answers have been recorded as (b - d), but unexpected correct answers have been recorded as (d - b). This is because when the response is incorrect, i.e., x = 0, then the index of unexpectedness is exp (b - d), but when the response is correct, i.e., x = 1, then the index is exp (d - b).

The earmark of unexpectedness in Table 4.3.2 is a positive difference, whether from (b - d) or (d - b). Corresponding values for z^2 can be looked up in Table 4.3.3 which gives either values of $z_0^2 = \exp(b - d)$ for unexpected incorrect answers or values of $z_1^2 = \exp(d - b)$ for unexpected correct answers. The entry C_x in Column 1 of Table 4.3.3 is either $C_0 = (b - d)$ when x = 0 and the response is incorrect or $C_1 = (d - b)$ when x = 1 and the response is correct.

TABLE 4.3.3

MISFIT STATISTICS

1 DIFFERENCE BETWEEN PERSON ABILITY AND ITEM DIFFICULTY C*	2 SQUARED STANDARDIZED RESIDUAL $z^2 = \exp C$	3 IMPROBABILITY OF THE RESPONSE $p = 1/(1 + z^2)$	4 RELATIVE EFFICIENCY OF THE OBSERVATION $I = 400p(1 - p)$	5 NUMBER OF ITEMS NEEDED TO MAINTAIN EQUAL PRECISION $L = 1000/I$
−0.6,0.4	1	.50	100	10
0.5,0.9	2	.33	90	11
1.0,1.2	3	.25	75	13
1.3,1.5	4	.20	65	15
1.6,1.7	5	.17	55	18
1.8,1.8	6	.14	50	20
1.9,2.0	7	.12	45	22
2.1	8	.11	40	25
2.2	9	.10	36	28
2.3	10	.09	33	30
2.4	11	.08	31	32
2.5	12	.08	28	36
2.6	13	.07	25	40
2.7	15	.06	23	43
2.8	16	.06	21	48
2.9	18	.05	20	50
3.0	20	.05	18	55
3.1	22	.04	16	61
3.2	25	.04	15	66
3.3	27	.04	14	73
3.4	30	.03	12	83
3.5	33	.03	11	91
3.6	37	.03	10	100
3.7	40	.02	9	106
3.8	45	.02	9	117
3.9	49	.02	8	129
4.0	55	.02	7	142
4.1	60	.02	6	156
4.2	67	.02	6	172
4.3	74	.01	5	189
4.4	81	.01	5	209
4.5	90	.01	4	230
4.6	99	.01	4	254

*For incorrect responses when x = 0 then $C_0 = (b - d)$. For correct responses when x = 1 then $C_1 = (d - b)$.

TABLE 4.3.4

FIT MEAN SQUARES (z^2_{vi})
FOR UNEXPECTED RESPONSES

	ITEM							
PERSON	4	5	7	6	8	12	14	PERSON MISFIT TOTAL
11	15							15
12				6	2			8
17		15						15
3						6		6
13			27	18		6		51
29			27		7		17	51
Item Misfit Total	15	15	54	24	9	12	17	146
		"1" expected "0" unexpected				"0" expected "1" unexpected		

We can locate the difference +2.7 for the (b - d) of Person 11 on Item 4 in the first column of Table 4.3.3 and read the corresponding z^2 in Column 2 as 15. This value and all of the other values for the differences in Table 4.3.2 have been recorded in Table 4.3.4, which now contains all the z^2 for every instance of unexpectedness that we have observed for the six persons and seven items. In the margins of Table 4.3.4 are the sums of these z^2 for each person and item. These sums indicate how unexpected the person or item pattern of responses is.

In Column 3 of Table 4.3.3 we show $p = 1/(1+z^2)$, the improbability of the observed response. This provides a significance level for the null hypothesis of fit for any particular response. With our example of a (b-d) of 2.7 we have a significance level of .06 against the null hypothesis that the response of Person 11 to Item 4 is according to the model. The z^2 themselves, are approximately χ^2 distributed with almost 1 degree of freedom each. When they are accumulated over items for a person or over persons for an item, the resulting sums are approximately χ^2 distributed with (L-1) degrees of freedom for a person and (N-1) degrees of freedom for an item.

In Column 4 of Table 4.3.3 we show $I = 400p(1-p)$, an index of the relative efficiency with which an observation at that (b-d) provides information about the person and item interaction. This index is scaled by the factor 400 so that it will give the amount

of information provided by the observation as a percentage of the maximum information that one observation at (b - d) = 0, i.e., right on target, could provide. The percent information in an observation can be used to judge the value of any particular item for measuring a person. This can be done by considering how much information would be lost by removing that item from the test. Thus, the I of 23% for Person 11 on Item 4 gives us an indication of how much we gain by including Item 4 in the measurement of Person 11 or of how much we would lose were we to remove Item 4.

The way the idea of information or efficiency enters into judging the value of an observation is through its bearing on the precision of measurement. Measurement precision depends on the number of items in the record and on the relevance of each item to the particular person. We can simplify the evaluation of each item's contribution to our knowledge of the person by calculating what percent of a best possible item the item in question contributes. That is what the values of I in Column 4 provide.

When the item and person are close to one another, i.e., on target, then the item contributes more to the measure of the person than when the item and person are far apart. The greater the difference between item and person, the greater the number of items needed to obtain a measure of comparable precision and, as a result, the less efficient each item.

For example, it requires five 20% items to provide as much information about a person as could be provided by one 100% item. Thus, when $(b-d)$ is about 3.0, it takes four to five times as many items to provide as much information as could be had from items that fell within one logit of the person, i.e., in the $|b-d| < 1$ region.

In general, the test length necessary to maintain a specified level of measurement precision is inversely proportional to the relative efficiency of the items used. The number L of less efficient items necessary to match the precision of 10 right-on-target items is given in the last column of Table 4.3.3.

To facilitate the use of Table 4.3.3, it has been arranged in four sections:

Right $\|b-d\| < 2$ on target	Item efficiency is 45% or better, in the $\|b-d\| < 1$ region, 79% or better. Misfit is difficult to detect.
Slightly $2 < \|b-d\| < 3$ off target	Efficiency is poor, less than 45%. Misfit becomes detectable when unexpected responses accumulate.
Rather $3 < \|b-d\| < 4$ off target	Efficiency is very poor, less than 18%. Even single unexpected responses can indicate significant response irregularities.
Extremely $4 < \|b-d\|$ off target	Efficiency is virtually nil, less than 7%. Unexpected responses are always unacceptable.

4.4 MISFITTING PERSON RECORDS

Upon examining the rows of Table 4.3.4 for high z^2 values in person records, we find that the highest accumulated values are for Persons 13 and 29. These are the two persons whose test behavior is most questionable, and so we will examine their records in more detail.

TABLE 4.4.1
COMPLETE FIT ANALYSIS FOR PERSONS 29 AND 13

		ITEM																	
	DIFFICULTY d_i		4 −3.9	5 −3.3	7 −3.3	6 −2.9	9 −2.9	8 −2.0	10 −1.4	1 0.6	13 1.5	12 1.7	14 2.8	15 4.3	16 4.3	17 4.3	SCORE r	SUM OF SQUARES Σz^2
PERSON 29 (b = 0.0)		x	1	1	0	1	1	0	1	1	0	0	1	0	0	0	7	
	x = 0	(b - d)			3.3			2.0			−1.5	−1.7		−4.3	−4.3	−4.3		
	x = 1	(d - b)	−3.9	−3.3		−2.9	−2.9		−1.4	0.6			2.8					
		z^2	0	0	27	0	0	7	0	2	0	0	17	0	0	0		53
PERSON 13 (b = 0.0)		x	1	1	0	0	1	1	1	1	0	1	0	0	0	0	7	
	x = 0	(b - d)			3.3	2.9					−1.5		−2.8	−4.3	−4.3	−4.3		
	x = 1	(d - b)	−3.9	−3.3			−2.9	−2.0	−1.4	0.6		1.7						
		z^2	0	0	27	18	0	0	0	2	0	6	0	0	0	0		53

Table 4.4.1 displays the response vectors for Persons 13 and 29 over all 14 items. For each person we show their responses of 0 or 1, the concomitant (b-d) or (d-b) differences, depending upon whether the response is 0 for incorrect or 1 for correct, and the consequent value of z^2. The sums of the row of z^2 for Person 13 and Person 29 are, coincidentally, 53. According to the model, these accumulated z^2's ought to follow a chi-square distribution with 1 degree of freedom for each z^2 minus the degree of freedom necessary to estimate the person measure b.

Further, any sum of z^2's, when divided by its degrees of freedom, should follow a mean square or $v = \Sigma z^2/f$ distribution which can conveniently be evaluated as the t — statistic:

$$v = \Sigma z^2/f \qquad \text{and} \qquad f = L - 1 \qquad [4.4.1]$$

$$t = [\ell n(v) + v - 1]\,[f/8]^{1/2} \sim N(0,1) \qquad [4.4.2]$$

which has approximately a unit normal distribution.

For Person 13 we have

$$v_{13} = \sum_{i}^{14} z^2_{13i}/(14 - 1) = 53/13 = 4.1$$

for which

$$t_{13} = [\ell n(v_{13}) + v_{13} - 1]\,[^{13}/_{8}]^{1/2} = [1.4 + 4.1 - 1]\,[1.3] = 5.8,$$

which is a rather improbable value for t, if this person's performance fits the model.

For Person 29 we observe the same results and the same t - statistic. With such significant misfit it would seem reasonable to diagnose these two records as unsuitable data sources either for the measurement of these two persons or for the calibration of these items.

4.5 MISFITTING ITEM RECORDS

We can also see in Table 4.3.4 that Items 7 and 6 show the greatest misfit among items, especially Item 7 with an accumulated z^2 of 54. In Table 4.5.1 we analyze the complete data vectors of these two items, showing for each person's response of 0 or 1 the associated (b-d) or (d-b) with their respective z^2.

For Item 7

$$v_7 = \sum_{\nu}^{34} z^2_{\nu,7}/(34-1) = 57/33 = 1.7$$

for which

$$t_7 = [\ell n(v_7) + v_7 - 1]\,[^{33}/_{8}]^{1/2} = [0.5 + 1.7 - 1]\,[2.0] = 2.4,$$

which is also a somewhat improbable value for t, if this item fits the model.

TABLE 4.5.1

COMPLETE FIT ANALYSIS FOR ITEM 7 AND 6

		ITEM 7 (d = – 3.3)				ITEM 6 (d = – 2.9)			
PERSON	ABILITY b	RESPONSE x	x = 0 (b - d)	x = 1 (d - b)	z^2	RESPONSE x	x = 0 (b - d)	x = 1 (d - b)	z^2
25	– 3.8	0	– 0.5		1	1		0.9	3
4	– 2.8	1		– 0.5	1	0	+ 0.1		1
33	– 2.8	1		– 0.5	1	0	+ 0.1		1
1	– 1.9	1		– 1.4	0	1		– 1.0	0
27	– 1.9	1		– 1.4	0	1		– 1.0	0
11	– 1.2	1		– 2.1	0	1		– 1.7	0
12	– 1.2	1		– 2.1	0	0	+ 1.7		6
17	– 0.6	1		– 2.7	0	1		– 2.3	0
19	– 0.6	1		– 2.7	0	1		– 2.3	0
30	– 0.6	1		– 2.7	0	1		– 2.3	0
2	0.0	1		– 3.3	0	1		– 2.9	0
3	0.0	1		– 3.3	0	1		– 2.9	0
5	0.0	1		– 3.3	0	1		– 2.9	0
6	0.0	1		– 3.3	0	1		– 2.9	0
8	0.0	1		– 3.3	0	1		– 2.9	0
9	0.0	1		– 3.3	0	1		– 2.9	0
(13)	0.0	0	+ 3.3		27	0	+ 2.9		18
16	0.0	1		– 3.3	0	1		– 2.9	0
26	0.0	1		– 3.3	0	1		– 2.9	0
28	0.0	1		– 3.3	0	1		– 2.9	0
(29)	0.0	0	+ 3.3		27	1		– 2.9	0
31	0.0	1		– 3.3	0	1		– 2.9	0
10	+ 0.6	1		– 3.9	0	1		– 3.5	0
18	+ 0.6	1		– 3.9	0	1		– 3.5	0
14	+ 0.6	1		– 3.9	0	1		– 3.5	0
32	+ 0.6	1		– 3.9	0	1		– 3.5	0
20	+ 0.6	1		– 3.9	0	1		– 3.5	0
21	+ 1.2	1		– 4.5	0	1		– 4.1	0
22	+ 1.2	1		– 4.5	0	1		– 4.1	0
23	+ 1.2	1		– 4.5	0	1		– 4.1	0
34	+ 1.2	1		– 4.5	0	1		– 4.1	0
15	+ 1.9	1		– 5.2	0	1		– 4.8	0
7	+ 2.8	1		– 6.1	0	1		– 5.7	0
24	+ 2.8	1		– 6.1	0	1		– 5.7	0
SUM OF SQUARES					57				29

For Item 6

$$v_6 = \sum_{\nu}^{34} z^2_{\nu 6} / (34\text{-}1) = 29/33 = 0.9$$

for which

$$t_6 = [\ell n(v_6) + v_6 - 1][^{33}/_8]^{1/2} = [-0.1 + 0.9 - 1][2.0] = 0.4,$$

obviously not a significant misfit.

We find that the mean square for Item 7 is significant but that the mean square for Item 6 is not. However, when we examine Table 4.5.1 again, we see that it is the two significantly misfitting persons 13 and 29 who contribute most to the misfit values for these two items. Now we have the opportunity of improving the fit of the data to the model, either by removing Item 7 and observing what happens then or by removing Persons 13 and 29.

4.6 BRIEF SUMMARY OF THE ANALYSIS OF FIT

For any response of Person v to Item i

$x_{\nu i} = 0$ if "incorrect" and

$x_{\nu i} = 1$ if "correct."

The standard mean square residual becomes

$$z^2_{\nu i} = \exp(b\text{-}d), \text{ for } x_{\nu i} = 0, \text{ incorrect, and}$$

$$z^2_{\nu i} = \exp(d\text{-}b), \text{ for } x_{\nu i} = 1, \text{ correct.}$$

To evaluate the overall fit of person v, we sum his vector of standard square residuals ($z^2_{\nu i}$) over the test of i = 1,L items, and calculate his person misfit statistic as

$$v_\nu = \sum_{i}^{L} z^2_{\nu i} / (L - 1) \sim F_{L-1,\,\infty} \qquad [4.6.1]$$

with

$$t_\nu = [\ell n(v_\nu) + v_\nu - 1][(L - 1)/8]^{1/2} \sim N(0,1) \qquad [4.6.2]$$

To evaluate the fit of Item i, we sum the item's vector of standard square residuals ($z^2_{\nu i}$) over the sample of v = 1,N persons, and calculate the item misfit statistic as

$$v_i = \sum_{\nu}^{N} z^2_{\nu i} / (N - 1) \sim F_{N-1,\,\infty}$$

with $t_i = [\ell n(v_i) + v_i - 1]\,[(N-1)/8]^{1/2} \sim N(0,1)$ **[4.6.4]**

4.7 COMPUTER ANALYSIS OF FIT

In the analysis of fit done by hand we saw that certain person records and items had residuals evaluated as significant. Having shown the procedures for the analysis of fit by hand we turn to computer analysis and return to our calibration of the KCT with 18 items and 34 persons. In the calibration of the KCT we see from the fit mean square, given in the left panel of Table 4.7.1, that Item 7 produces the greatest misfit with a value of 1.98 not far from the 1.7 found in our hand computation. From our analysis of person misfit we know that Persons 13 and 29 greatly contributed to this misfit in Item 7. Without this information at the time of our calibration, however, we might have considered the possible deletion of Item 7 because of its high fit mean square. With this much lack of fit for Item 7 we might have chosen to recalibrate with Item 7 removed. This has been done and the results are given in the middle panel of Table 4.7.1. Now we see that Item 6 has acquired a misfit of 2.73 even though previously when we calibrated all 14 items, Item 6 had a fit mean square of only 0.90. This change in the status of Item 6 is troublesome. We do not seem to be focusing in on a set of suitable items. Nevertheless we go one step further and recalibrate once more, this time removing both Item 7 and Item 6. The results are in the right panel of Table 4.7.1. Alas, now we find that Item 8 has become a misfit. These attempts to find a properly fitting set of items appear doomed.

TABLE 4.7.1
ANALYSIS OF FIT WITH UCON : ITEM DELETIONS

ITEMS IN FIT ORDER				ITEMS IN FIT ORDER				ITEMS IN FIT ORDER			
SEQ NUM	ITEM NAME	ITEM DIFF	FIT MN SQ	SEQ NUM	ITEM NAME	ITEM DIFF	FIT MN SQ	SEQ NUM	ITEM NAME	ITEM DIFF	FIT MN SQ
16	6	4.56	0.13								
17	6	4.56	0.13	16	6	4.45	0.13				
15	6	4.56	0.13	17	6	4.45	0.13	16	6	4.29	0.13
9	4	- 3.22	0.23	15	6	4.45	0.13	17	6	4.29	0.13
4	3	- 4.19	0.37	9	4	- 3.75	0.22	15	6	4.29	0.13
13	5	1.86	0.40	4	3	- 4.75	0.40	9	4	- 4.30	0.18
8	4	- 2.24	0.44	13	5	1.77	0.42	4	3	- 5.38	0.35
5	3	- 3.65	0.53	5	3	- 4.20	0.54	13	5	1.61	0.44
11	5	0.76	0.77	11	5	0.65	0.68	5	3	- 4.79	0.64
10	4	- 1.50	0.79	14	6	3.11	0.74	11	5	0.43	0.67
6	3	- 3.22	0.90	12	5	2.04	0.83	14	6	2.96	0.77
12	5	2.14	0.97	10	4	- 1.81	0.88	10	4	- 2.19	0.78
14	6	3.21	1.33	8	4	- 2.66	1.03	12	5	1.89	0.82
[7]	4	- 3.65	[1.98]	[6]	3	- 3.75	[2.73]	[8]	4	- 3.11	[1.29]
All Persons and All Items				Deleting Item 7				Deleting Items 7 and 6			
L = 14 N = 34				L = 13 N = 34				L = 12 N = 34			

Suppose, instead, we decide, subsequent to our first calibration of the KCT items, to evaluate person fit. The computer analysis for person misfit, shown in Table 4.7.2, also identifies Person 13 and 29 as producing the highest fit statistics. So let us recalibrate all 14 of the items but with these two persons removed. Now, in Table 4.7.3, we see that the fit mean squares for all of the items are small enough to satisfy us. Removing the two unsuitable person records has brought all of the items into agreement.

TABLE 4.7.2

ANALYSIS OF PERSON FIT WITH UCON

PERSON	SCORE r	UCON ABILITY b	UCON MISFIT v	
25	2	− 4.4	0.5	
4	3	− 3.7	0.4	
33	3	− 3.7	0.9	
1	4	− 3.1	0.3	
27	4	− 3.1	0.3	
11	5	− 2.3	0.8	
12	5	− 2.3	0.5	
17	6	− 1.4	1.0	
19	6	− 1.4	0.2	
30	6	− 1.4	0.2	
2	7	− 0.3	0.1	
3	7	− 0.3	1.4	
5	7	− 0.3	0.1	
6	7	− 0.3	0.1	
8	7	− 0.3	0.1	
9	7	− 0.3	0.1	
(13)	7	− 0.3	[5.7]	(Hand PROX = 4.1)
16	7	− 0.3	0.6	
26	7	− 0.3	0.1	
28	7	− 0.3	0.6	
(29)	7	− 0.3	[6.6]	(Hand PROX = 4.1)
31	7	− 0.3	0.1	
10	8	+ 1.0	0.2	
14	8	+ 1.0	0.2	
18	8	+ 1.0	0.4	
20	8	+ 1.0	0.4	
32	8	+ 1.0	0.2	
21	9	+ 2.0	0.2	
22	9	+ 2.0	0.2	
23	9	+ 2.0	0.7	
34	9	+ 2.0	0.7	
15	10	+ 3.0	0.2	
7	11	+ 3.9	0.4	
24	11	+ 3.9	0.9	
		Mean	0.7	
		Standard Deviation	1.6	

TABLE 4.7.3

ANALYSIS OF FIT WITH UCON : PERSON DELETIONS

SEQ NUM	ITEM NAME	ITEM DIFF	FIT MN SQ
7	4	– 5.70	0.10
16	6	5.27	0.13
17	6	5.27	0.13
15	6	5.27	0.13
8	4	– 2.87	0.17
9	4	– 3.73	0.21
6	3	– 4.24	0.34
13	5	2.34	0.38
4	3	– 4.84	0.40
14	6	4.43	0.55
5	3	– 4.24	0.64
11	5	1.51	0.70
12	5	3.01	0.99
10	4	– 1.48	1.03

Deleting Persons 13 and 29

L = 14 N = 32

It seems clear that it was the test records of these two unpredictable persons which caused Item 7 and then Item 6 to seem to misfit. Thus, we learn that successive deletions of items without analyzing person fit can lead us to believe that items are misfitting when, in fact, it is the response records of a few irregular persons which are causing the trouble. While the very small sample size used in our example exaggerates the impact of the two irregular persons, even large samples do not completely obliterate the contaminating influence of irregular person records, and in a large sample such flawed records may be harder to spot and so remain unknown unless explicit tests of person fit are routinely made.

5 CONSTRUCTING A VARIABLE

5.1 GENERALIZING THE DEFINITION OF A VARIABLE

In Chapters 2, 3 and 4 we have shown how to expose and evaluate the observed relationship between intended measuring instruments, the test items, and the objects they are intended to measure, the persons. This prepares us for the present chapter which is concerned with how to define a variable.

With a workable calibration procedure and a method for the evaluation of fit, it becomes practical to turn our attention to a far more important activity, namely a critical examination of the calibrated items to see what it is that they imply about the possibility of a variable of some useful generality. We want to find out whether our calibrated items spread out in a way that shows a coherent and meaningful direction. If they are not spread out at all, then all we have achieved is to define a point, perhaps on some variable, perhaps not. But the variable itself, whatever it may be remains obscure.

Our intention now is to show how calibrated items can be used to define a variable and how to find out whether the resulting operational definition of the variable makes sense. We will begin by examining the degree to which the spread of item difficulties substantially exceeds the standard error of their estimates, that is, the degree to which the data has given a direction to the variable. For example, suppose we consider the estimates of two item difficulties with their respective standard errors. In order for these two items to define a line between them the difference between their estimates must be substantially greater than the standard error of this difference! Only if the two estimates are well separated by several such standard errors will we begin to see a line between the two items suggesting a direction for the variable which they define.

If, however, when we compare these two estimates by a standard error or two, they overlap substantially, then we cannot assume that the two values differ and as a result no direction for a variable has been defined. Instead the items define a point without direction.

Figure 5.1.1 illustrates this. In Example 1 we have Items A and B separated from each other by several standard errors. Even with two items we begin to see a direction to the variable at least as defined by these two items. In the second example, however, we find the two items so close to each other that, considering their standard errors, they are not separable. We have found a point. But no direction has been established and so no variable has as yet been implied.

As an example of variable definition, we will continue our study of the KCT data to see how well the KCT items succeed in defining a variable and just what that variable seems to be.

5.2 DEFINING THE KCT VARIABLE

The items of the KCT form a tapping series that grows in length by increasing the number of taps and grows in complexity by the distance between adjacent taps and the number of reverses in direction of movement.

FIGURE 5.1.1

DEFINING A VARIABLE

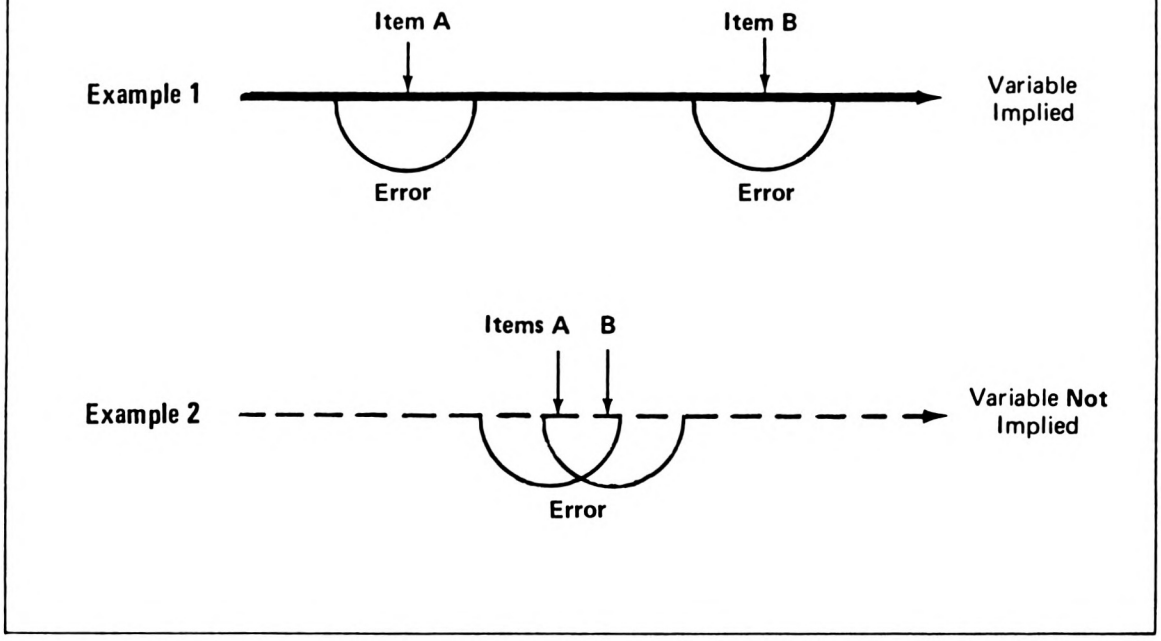


Figure 5.2.1 lists the 18 items comprising the original KCT. Each item is described by its numerical name, tapping series and tapping order pattern.

Table 5.2.1 focuses on those 14 KCT items that were calibrated in Chapters 2 and 3. Items 1, 2 and 3 are not included because they were too easy for the 34 persons in that sample and Item 18 is not included because it was too hard. Table 5.2.1 gives the item names, tapping series, item difficulties and their standard errors. The difficulty range of these 14 items is from -4.2 logits to +4.6 logits.

The item difficulties in Table 5.2.1 make it possible to be quantitatively explicit in our definition of the KCT variable by placing the 14 items at their calibrated positions along the line of the variable. This is done in Figure 5.2.2 As several items have either the same difficulty or are so close in terms of their standard errors that they can hardly be differentiated, we have shown only the eight items that best mark out the extent of the KCT variable. The semicircles in Column 2 of Figure 5.2.2 show an allowance of one standard error around each estimated difficulty. We can see that Items 4, 6, 8 and 10 define the easy end of the variable. Then there is a rather wide undefined gap in the middle. Finally, Items 11, 12, 14 and 16 define the hard end. The tapping patterns in Column 3 show what movement along the variable means in terms of the increasing number of taps and pattern complexity. Column 4 gives the distribution along this KCT variable of the 34 persons who participated in the initial calibration.

FIGURE 5.2.1

DESCRIPTION OF THE KCT VARIABLE

ITEM NAME	TAPPING SERIES	BLOCK NUMBERS	TAPPING PATTERN
1	1 – 4	4 3 2 1	
2	2 – 3	4 3 2 1	
3	1 – 2 – 4	4 3 2 1	
4	1 – 3 – 4	4 3 2 1	
5	2 – 1 – 4	4 3 2 1	
6	3 – 4 – 1	4 3 2 1	
7	1 – 4 – 3 – 2	4 3 2 1	
8	1 – 4 – 2 – 3	4 3 2 1	
9	1 – 3 – 2 – 4	4 3 2 1	

FIGURE 5.2.1

DESCRIPTION OF THE KCT VARIABLE
(Continued)

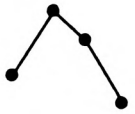
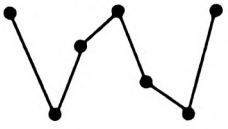
ITEM NAME	TAPPING SERIES	BLOCK NUMBERS	TAPPING PATTERN
10	2 – 4 – 3 – 1	4 3 2 1	
11	1 – 3 – 1 – 2 – 4	4 3 2 1	
12	1 – 3 – 2 – 4 – 3	4 3 2 1	
13	1 – 4 – 3 – 2 – 4	4 3 2 1	
14	1 – 4 – 2 – 3 – 4 – 1	4 3 2 1	
15	1 – 3 – 2 – 4 – 1 – 3	4 3 2 1	
16	1 – 4 – 2 – 3 – 1 – 4	4 3 2 1	
17	1 – 4 – 3 – 1 – 2 – 4	4 3 2 1	
18	4 – 1 – 3 – 4 – 2 – 1 – 4	4 3 2 1	

TABLE 5.2.1

CALIBRATION OF THE KCT VARIABLE WITH ITEMS IN ORDER OF DIFFICULTY

ITEM NAME	TAPPING SERIES	ITEM CALIBRATION	STANDARD ERROR
4	1 – 3 – 4	-4.2	0.8
5	2 – 1 – 4	-3.6	0.7
7	1 – 4 – 3 – 2	-3.6	0.7
6	3 – 4 – 1	-3.2	0.6
9	1 – 3 – 2 – 4	-3.2	0.6
8	1 – 4 – 2 – 3	-2.2	0.6
10	2 – 4 – 3 – 1	-1.5	0.5
11	1 – 3 – 1 – 2 – 4	0.8	0.5
13	1 – 4 – 3 – 2 – 4	1.9	0.5
12	1 – 3 – 2 – 4 – 3	2.1	0.6
14	1 – 4 – 2 – 3 – 4 – 1	3.2	0.7
15	1 – 3 – 2 – 4 – 1 – 3	4.6	1.1
16	1 – 4 – 2 – 3 – 1 – 4	4.6	1.1
17	1 – 4 – 3 – 1 – 2 – 4	4.6	1.1
	Mean	0.0	
	Standard Deviation	3.4	

UCON Calibration from Table 3.4.1

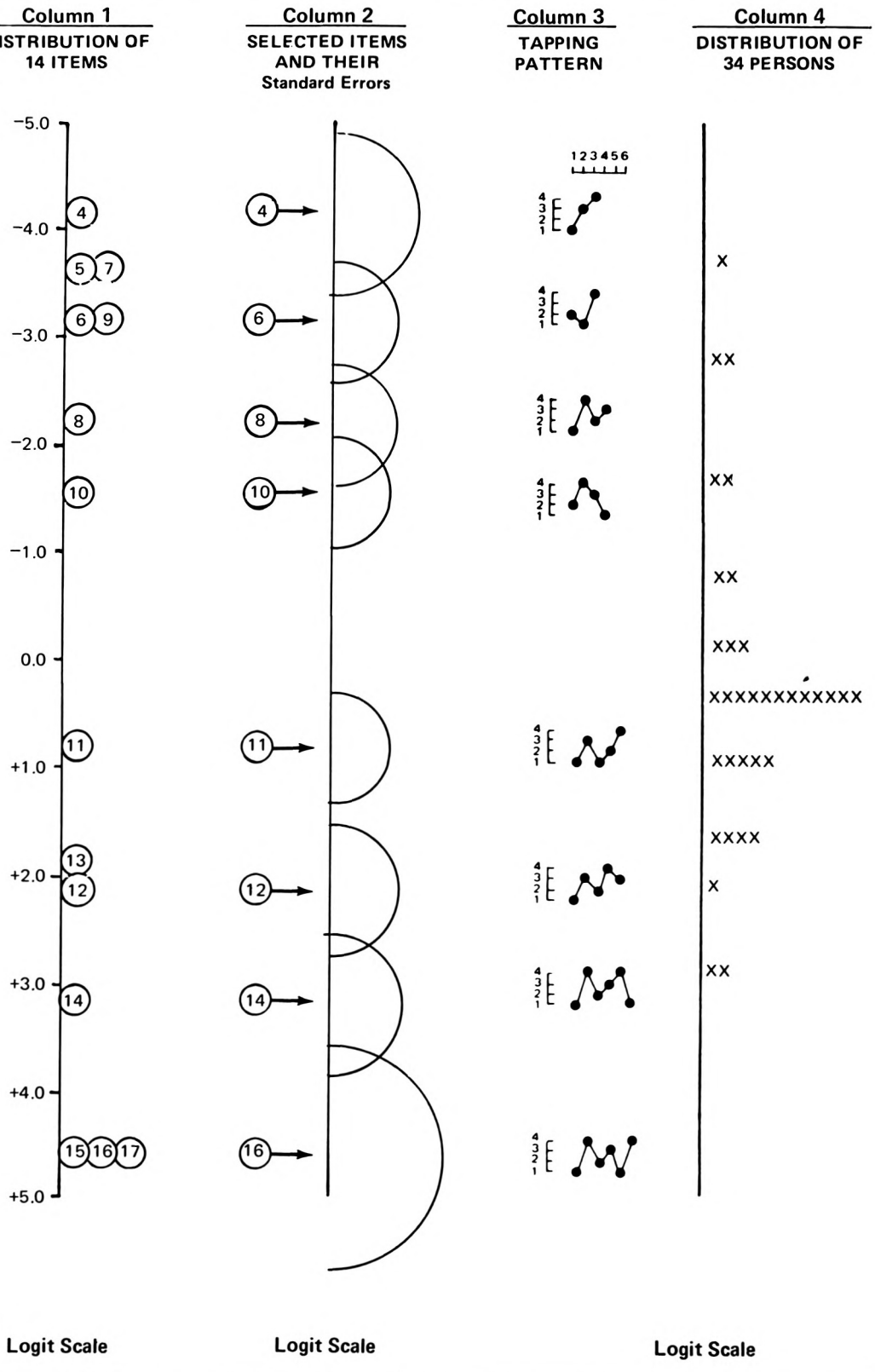
We see that most of the persons in this sample fall in the center of the test. But that is just where we have a large gap in test items. We have discovered something important and useful to us, namely that our test instrument is weakest at the mode of our sample. It becomes clear that, if we want to discriminate among the majority of persons found in the middle range of the KCT, then we must construct some additional middle range items which will be more appropriate to middle range abilities.

5.3 INTENSIFYING AND EXTENDING THE KCT VARIABLE

To improve measurement along the KCT variable, especially in the middle range, further item development is required. We need items to fill the gap in the original definition of the variable and we need easier and harder test items in order to extend the variable's range. However, since all of our sample passed the three easiest items, extending the KCT variable down to easier levels may prove difficult. We would have to locate some much less able persons than were found among our original 34 in order to calibrate easier items. On the other hand, only one hard item was failed by all persons in our KCT sample. It might be fruitful to try to add some items which are more difficult than Items 15, 16 and 17, under the assumption that with a sample of more able persons we could obtain useful calibrations of these more difficult items and thus extend the KCT variable upward.

FIGURE 5.2.2

DEFINING THE KCT VARIABLE BY ITEM DIFFICULTY DISTRIBUTION, TAPPING PATTERN AND PERSON ABILITY DISTRIBUTION



With these considerations in mind, further development of the KCT variable was undertaken. All 18 items from the original KCT were retained, and ten new items were added. The original KCT was from Form II of the Arthur Point Scale. We examined Form I and found three items not used in Form II (Arthur, 1943). To these three items we added seven more. Five items were designed to fill the middle range gap, four items were designed to extend the KCT variable upward and one of the Form I items was expected to fit near old Items 5, 6 and 7. The tapping series for these additional items and their intended locations on the KCT variable are shown in Figures 5.3.1 and 5.3.2.

Figure 5.3.1 shows the one item from Form I and the five new items designed to fill the gap between the old KCT Items 10 and 11. The four items designed to extend the KCT in the region of Item 18 are shown in Figure 5.3.2. The result is a new test form, KCTB, which contains all 18 old items and, in addition, 10 new items. This new instrument of 28 items was administered to a sample of 101 persons and Items 2 through 25 were calibrated. Item 1 was still too easy and Items 26, 27 and 28 were still too hard to be calibrated.

Column 6 in Table 5.3.1 gives these new KCTB calibrations. The rest of Table 5.3.1 shows the relationship between the old KCT and the new KCTB calibrations. Column 1 names the 14 old KCT items. Column 2 shows their original calibrations from Table 3.4.4. Notice in Column 6 that we have now obtained calibrations on old KCT Items 2, 3 and 18, three of the original items which remained uncalibrated in our first study with 34 persons.

Column 3 of Table 5.3.1 applies the necessary adjustment to bring the old KCT calibrations into line with their new calibrations on the new KCTB. This is done by shifting the calibrations in Column 2 by the constant 0.4 which is the mean position of the old KCT items in the new KCTB calibrations. This causes Column 3 and Column 5 to have the same mean of 0.4.

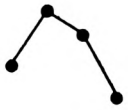
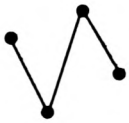
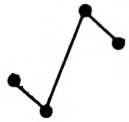
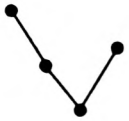
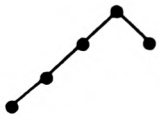
In Table 5.3.1 we see that the new KCTB Items 12 through 16 fall more or less where expected, if somewhat on the easy side. KCTB Item 25 along with KCT Item 18 extend the reach of the KCT variable 2 logits further upwards, but we have found no one who succeeds on KCTB Items 26, 27 and 28.

Figure 5.3.3 compares the difficulties of those items which appeared in both the KCT and KCTB calibrations. Each of the 14 items is located in Figure 5.3.3 by its pair of difficulty estimates. If the items fit the measurement model, then we expect these independent estimates of their difficulties to be statistically equivalent.

Thus the extent to which the 14 points fall along the identity line tests the invariance of these 14 items difficulties. As Figure 5.3.3 shows, the 14 points all lie well within 95% quality control lines. This is the pattern that the model says they must approximate in order to be useful as instruments of measurement.

FIGURE 5.3.1

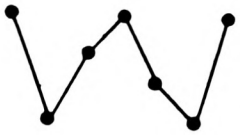
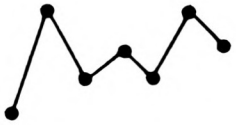
FIVE NEW ITEMS DESIGNED TO FILL THE GAP IN KCT AND ONE EASIER ITEM FROM KCT, FORM I

ITEM NAME Old KCT	New KCTB	TAPPING SERIES	BLOCK NUMBERS	TAPPING PATTERN
	7	1 – 2 – 3 – 4	4 3 2 1	
10	11	2 – 4 – 3 – 1	4 3 2 1	
	12	3 – 1 – 4 – 2	4 3 2 1	
	13	2 – 1 – 4 – 3	4 3 2 1	
	14	4 – 2 – 1 – 3	4 3 2 1	
	15	1 – 2 – 3 – 4 – 3	4 3 2 1	
	16	1 – 2 – 3 – 4 – 2	4 3 2 1	
11	17	1 – 3 – 1 – 2 – 4	4 3 2 1	

Old KCT $\begin{matrix}\text{Items 1 - 6}\\ \text{Items 7 - 9}\end{matrix}$ become New KCTB $\begin{matrix}\text{Items 1 - 6}\\ \text{Items 8 - 10}\end{matrix}$

FIGURE 5.3.2

FOUR NEW KCTB ITEMS DESIGNED TO EXTEND KCT TO MEASURE MORE ABLE PERSONS

ITEM NAME Old KCT	ITEM NAME New KCTB	TAPPING SERIES	BLOCK NUMBERS	TAPPING PATTERN
17	23	1 – 4 – 3 – 1 – 2 – 4	4 3 2 1	
18	24	4 – 1 – 3 – 4 – 2 – 1 – 4	4 3 2 1	
	25	3 – 2 – 4 – 1 – 3 – 4 – 2	4 3 2 1	
	26	2 – 4 – 3 – 1 – 4 – 2 – 3	4 3 2 1	
	27	1 – 4 – 2 – 3 – 2 – 4 – 3	4 3 2 1	
	28	1 – 3 – 4 – 2 – 4 – 2 – 3	4 3 2 1	

Old KCT Items 12 - 16 become New KCTB Items 18 - 22

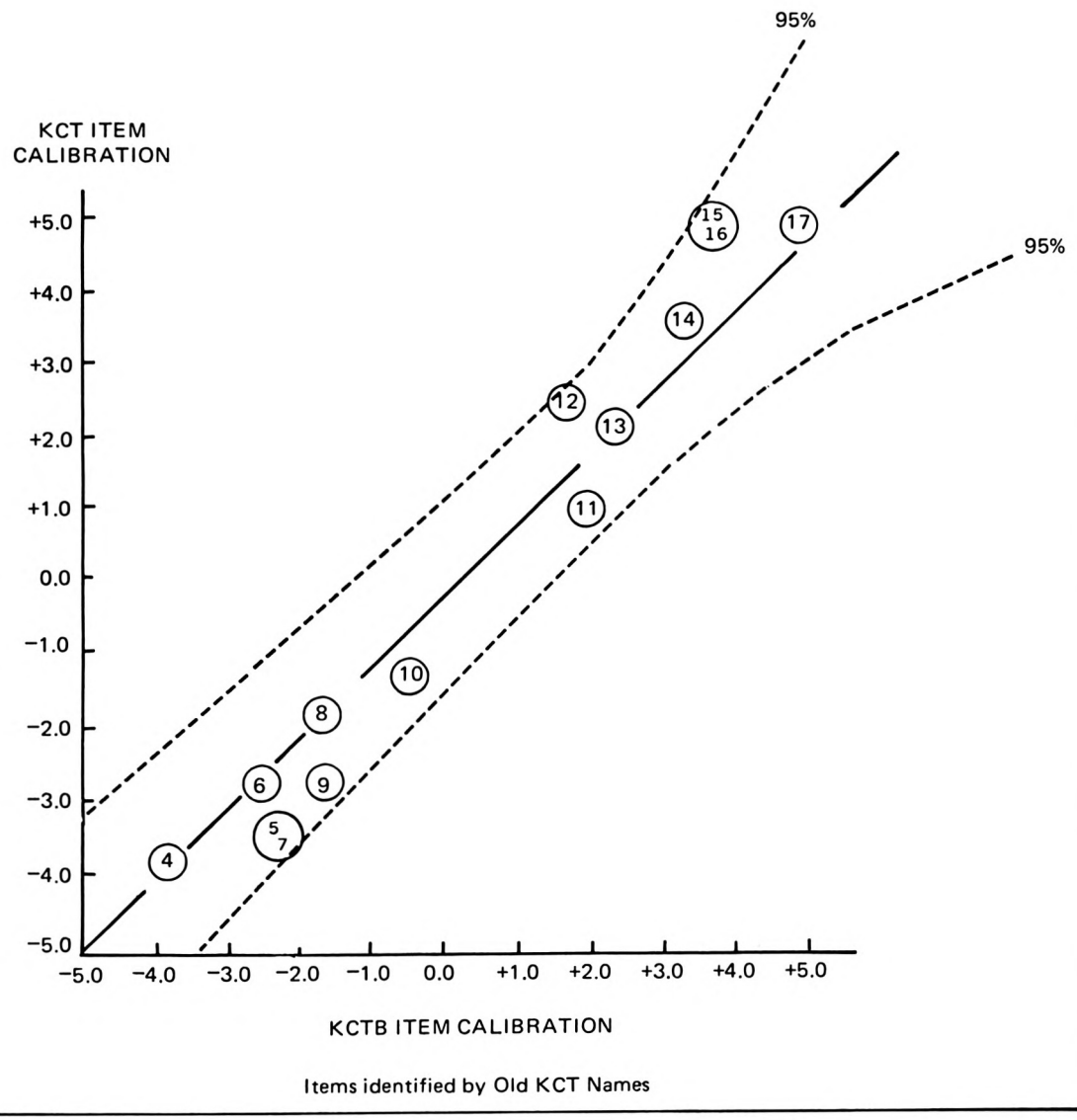
TABLE 5.3.1

CALIBRATION OF KCTB

1	2	3	4	5	6
	Old KCT			New KCTB	
Item	KCT Calibration		Item	KCTB Calibration	
Name	Unadjusted	Adjusted *	Name	Old Items	All Items
2			2		-6.0
3			3		-5.6
4	-4.2	-3.8	4	-3.8	-3.8
5	-3.6	-3.2	5	-2.3	-2.3
6	-3.2	-2.8	6	-2.5	-2.5
			7		-4.0
7	-3.6	-3.2	8	-2.3	-2.3
8	-2.2	-1.8	9	-1.8	-1.8
9	-3.2	-2.8	10	-1.8	-1.8
10	-1.5	-1.1	11	-0.8	-0.8
			12		0.1
			13		-0.6
			14		-0.3
			15		-1.3
			16		-0.5
11	0.8	1.2	17	2.2	2.2
12	2.1	2.5	18	1.6	1.6
13	1.9	2.3	19	2.2	2.2
14	3.2	3.6	20	3.1	3.1
15	4.6	5.0	21	3.6	3.6
16	4.6	5.0	22	3.6	3.6
17	4.6	5.0	23	4.7	4.7
18			24		6.5
			25		6.0
Mean	0.0	0.4		0.4	0.0
Standard Deviation	3.4	3.4		2.8	3.4
KCT:	L = 14	N = 34	KCTB:	L = 24	N = 101

*The Chapter 3 calibrations of the 14 old KCT items in Column 2 have been shifted along the variable by 0.4 logits so that the mean of these Chapter 3 calibrations equals their mean calibration in the new KCTB calibrations. This new mean was calculated from Column 5.

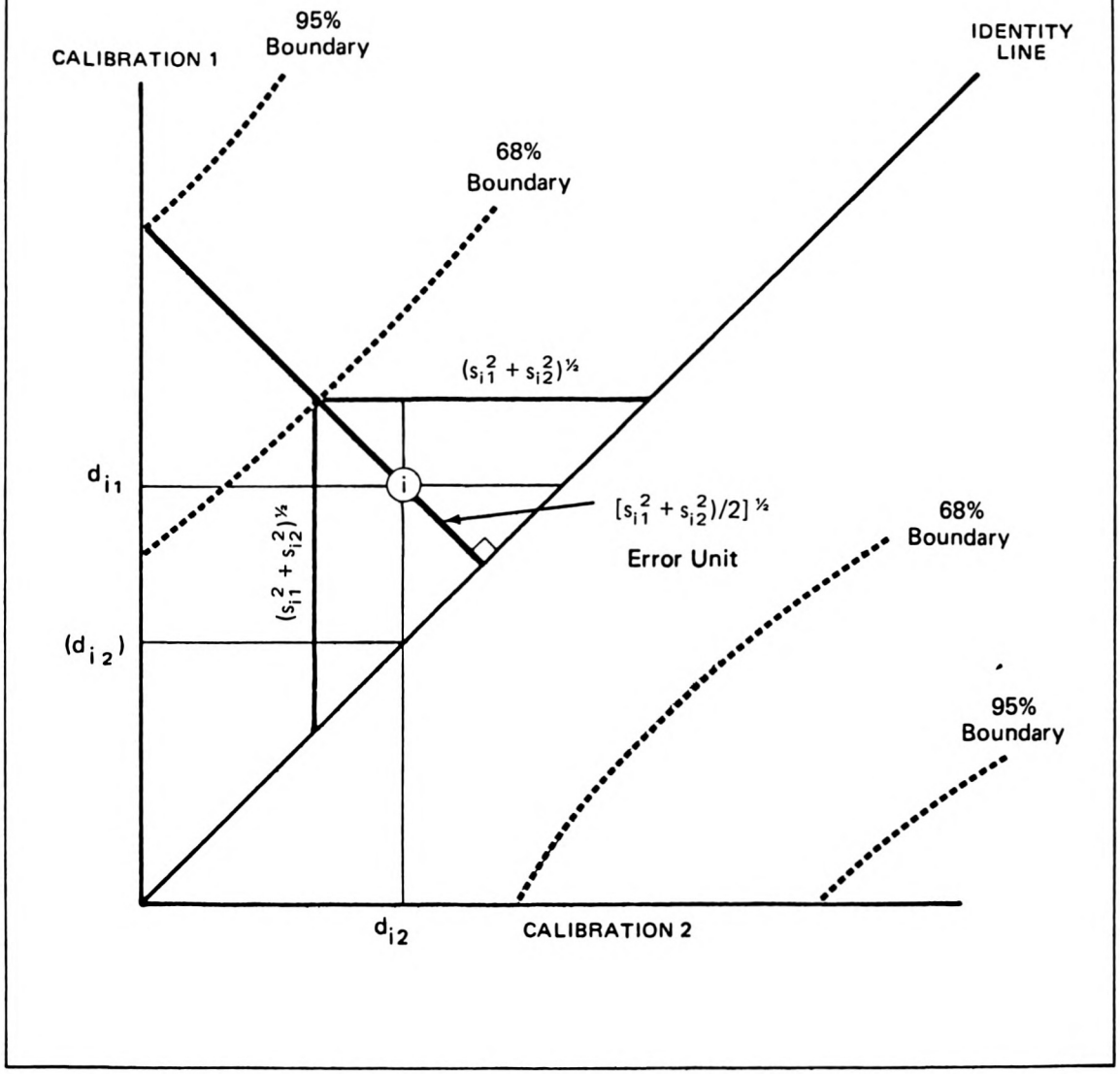
FIGURE 5.3.3
PLOT OF ITEM CALIBRATIONS,
KCT VERSUS KCTB



Items identified by Old KCT Names

FIGURE 5.4.1

HOW TO FORM 68% AND 95% QUALITY CONTROL LINES



5.4 CONTROL LINES FOR IDENTITY PLOTS

Figure 5.3.3 contains a pair of 95% quality control lines which help us see the extent to which the 14 item points conform to our model expectation of item difficulty invariance. In plots which are used to evaluate the invariance of item difficulty and hence the quality of items, these 95% lines make it easy to see how satisfactorily the item points in the plot follow the expected identity line.

Figure 5.4.1 shows how such lines are drawn. Each plot compares a series of paired item calibrations. Each item has a difficulty d_i and a standard error s_i from each of two independent calibrations in which the item appeared. Thus for each item i we have (d_{i1}, s_{i1}) and (d_{i2}, s_{i2}). Since each pair of calibrations applies to one item, we expect the two difficulties d_{i1} and d_{i2}, after a single translation necessary to establish an origin common to both sets of items, to estimate the same difficulty δ_i. We also expect the error of these estimates to be estimated by s_{i1} and s_{i2}.

This gives us a statistic for testing the extent to which the two d_i's estimate the same δ_i, namely

$$t_{i12} = (d_{i1} - d_{i2})/(s_{i1}^2 + s_{i2}^2)^{1/2} \qquad \sim N(0,1) \qquad [5.4.1]$$

in which $(s_{i1}^2 + s_{i2}^2)^{1/2}$ estimates the expected standard error of the difference between the two independent estimates d_{i1} and d_{i2} of the one parameter δ_i. We can introduce this test for the quality of each item point into the plot by drawing quality control boundaries at about two of these standard errors away from the identity line on each side.

Since the standard unit of difference error parallel to either axis of the plot is

$$(s_{i1}^2 + s_{i2}^2)^{1/2} ,$$

the unit of error perpendicular to the 45 degree identity line must be

$$[(s_{i1}^2 + s_{i2}^2)/2]^{1/2} .$$

Two of these error units perpendicular to the identity line in each direction yields a pair of approximately 95% quality control lines. The perpendicular distance $D_{i1\ 2}$ between these quality control lines and the identity line thus becomes

$$D_{i1\ 2} = 2[(s_{i1}^2 + s_{i2}^2)/2]^{1/2} . \qquad [5.4.2]$$

When s_{i1} and s_{i2} are sufficiently similar so that the mean of their squares is approximately the same as the square of their mean, that is

$$(s_{i1}^2 + s_{i2}^2)/2 \simeq [(s_{i2} + s_{i2})/2]^2 .$$

then the distance D_{i12} from the identity line to a 95% confidence boundary can be approximated by

$$D_{i12} = 2[(s_{i1}^2 + s_{i2}^2)/2]^{1/2}$$

$$\simeq 2[(s_{i1} + s_{i2})/2] = s_{i1} + s_{i2} .$$

Thus for the i = 1, K items for which paired calibrations are available the distances $(s_{i1} + s_{i2})$ perpendicular to the identity line drawn through each item point can be used to locate 95% confidence lines for evaluating the overall stability of the item calibrations shown in the plot.

5.5. CONNECTING TWO TESTS

The usual method for equating tests is based on the equation of equal-percentile scores. This procedure requires a sample of persons large enough and broadly enough distributed to assure an adequate definition of each score-to-percentile connection. With Rasch measurement a more economical and better controlled method for building an item bank becomes possible. Links of 10 to 20 common items can be embedded in pairs of otherwise different tests. Each test can then be administered to its own separate sample of persons. No person need take more than one test. But all items in all tests can be subsequently connected through the network of common item links.

To begin with a simple example, a traditional approach to equating two 60-item tests A and B might be to give them simultaneously to a sample of at least 1200 persons as depicted in the upper part of Figure 5.5.1. This is a likely plan since a detailed definition of score percentiles is necessary for successful percentile equating. Each person must take both tests, 120 items.

In contrast, a Rasch approach could do the same job with each person taking only one test of 60 items. To accomplish this a third 60-item test C is made up of 30 items from each of the original tests A and B. Then each of these three tests is given to a sample of 400 persons as depicted in the lower part of Figure 5.5.1. Now each person takes only one test, but all 120 items are calibrated together through the two 30-item links connecting the three tests. The testing burden on each person is one-half of that required by the equal-percentile plan.

In Rasch equating the separate calibrations of each test produce a pair of independent item difficulties for each linking item. According to the model, the estimates in each pair are statistically equivalent except for a single constant of translation common to all pairs in the link. If two tests, A and B, are joined by a common link of K items, each test is given to its own sample of N persons, and d_{iA} and d_{iB} are the estimated difficulties of item i in each test with standard errors of about $2.5/N^{1/2}$, then the single constant necessary to translate all item difficulties in the calibration of Test B onto the scale of Test A is

$$G_{AB} = \sum_{i}^{K} (d_{iA} - d_{iB})/K \qquad [5.5.1]$$

with a standard error of about $3.5/(NK)^{1/2}$ logits.

The quality of this link can be evaluated by the fit statistic

$$\sum_{i}^{K} (d_{iA} - d_{iB} - G_{AB})^2 (N/12)[K/(K-1)] \sim \chi^2_K \qquad [5.5.2]$$

which according to the model should be approximately chi-square with K degrees of freedom.

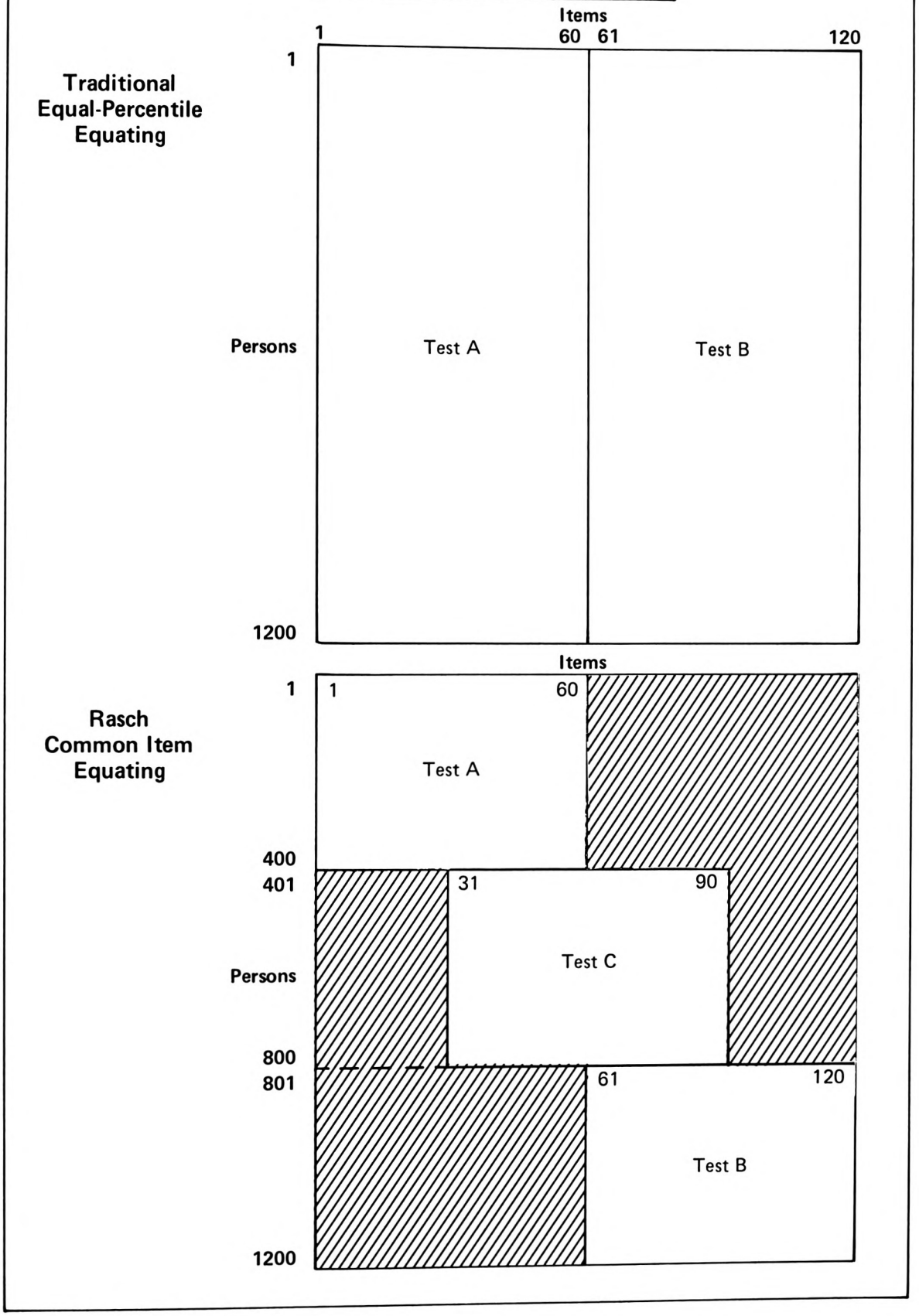
The individual fit of any item in the link can be evaluated by

$$(d_{iA} - d_{iB} - G_{AB})^2 (N/12)[K/(K-1)] \sim \chi^2_1 \qquad [5.5.3]$$

which according to the model should be approximately chi-square with one degree of freedom.

FIGURE 5.5.1

TRADITIONAL AND RASCH EQUATING DESIGNS



In using these chi-square statistics to judge link quality we must not forget how they are affected by sample size. When N exceeds 500 these chi-squares can detect link flaws too small to make any tangible difference in G_{AB}. When calibration samples are large the root mean square misfit is more useful. This statistic can be used to estimate the logit increase in calibration error caused by link flaws.

In deciding how to act on evaluations of link fit, we must also keep in mind that random uncertainty in item difficulty of less than .3 logits has no practical bearing on person measurement (Wright and Douglas, 1975a, 35-39). Because of the way sample size enters into the calculation of item difficulty and hence into the evaluation of link quality, we can deduce that samples of 200 persons and links of 10 good items will always be more than enough to supervise link validity at better than .3 logits. In practice we have found that we can construct useful item banks with sample units as small as 100 persons.

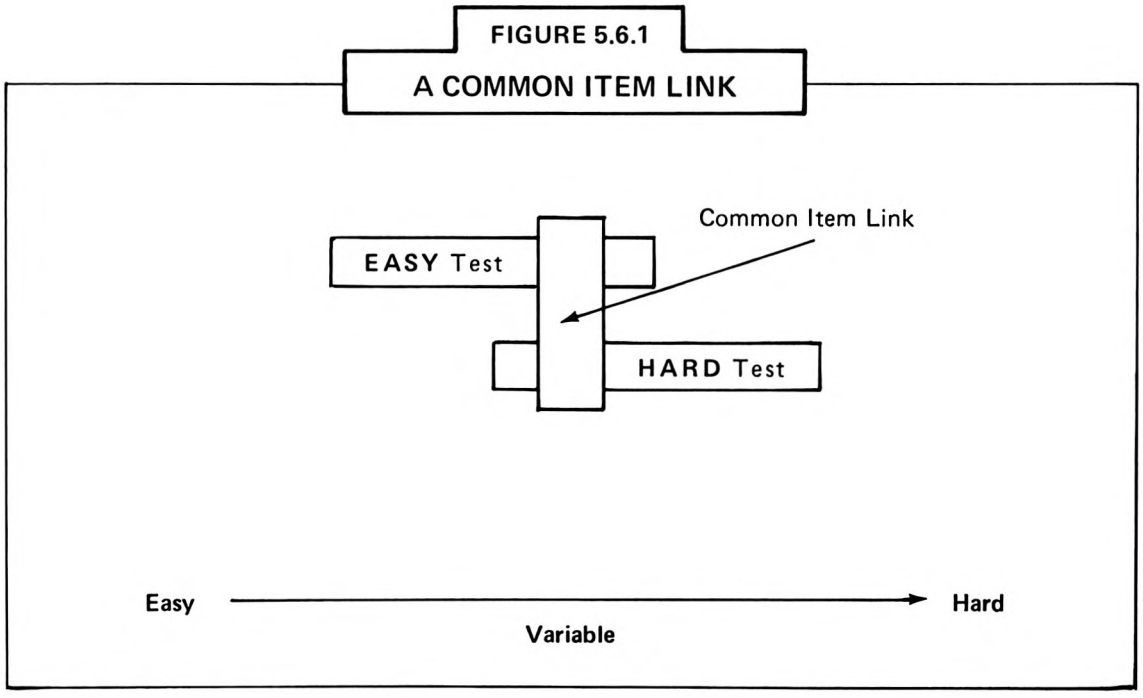
5.6 BUILDING ITEM BANKS

As we establish and extend the definition of a variable by the addition of new items we have the beginning of an item bank. With careful planning we can introduce additional items systematically and in this way build up a bank of calibrated items useful for an increasing variety of measurement applications. As the number of items increases, the problems of managing such a bank multiply. There is not only the question of how best to select and combine items and persons, but of how to manage effectively the consequent collection of calibrated items. Rasch measurement provides a specific well-defined approach to managing item banking.

The basic structure necessary to calibrate many items onto a single variable is the common item link in which one set of linking test items is shared by and so connects together two otherwise different tests. An easy and a hard test could be linked by a common set of items as pictured in Figure 5.6.1. In this example the linking items are the "hard" items in the EASY test but the "easy" items in the HARD test.

FIGURE 5.6.1

A COMMON ITEM LINK



With two or more test links we can build a chain of the kind shown in Figure 5.6.2. The representation in Figure 5.6.2, however, is awkward. The linking structure can be conveyed equally well by the simpler scheme in Figure 5.6.3 which emphasizes the links and facilitates diagramming more complicated structures.

As the number and difficulty range of the items introduced into an item bank grows beyond the test-taking capacity of any one person, the chain of items must be parceled into test forms of manageable length and difficulty range. In Figure 5.6.3 each circle indicates a test sufficiently narrow in range of item difficulties to be manageable by a suitably chosen sample of persons. Each line connecting a circle represents a link of common items shared by the two tests it joins. Tests increase in difficulty horizontally along the variable and are comparable in difficulty vertically.

FIGURE 5.6.2

A CHAIN WITH TWO LINKS

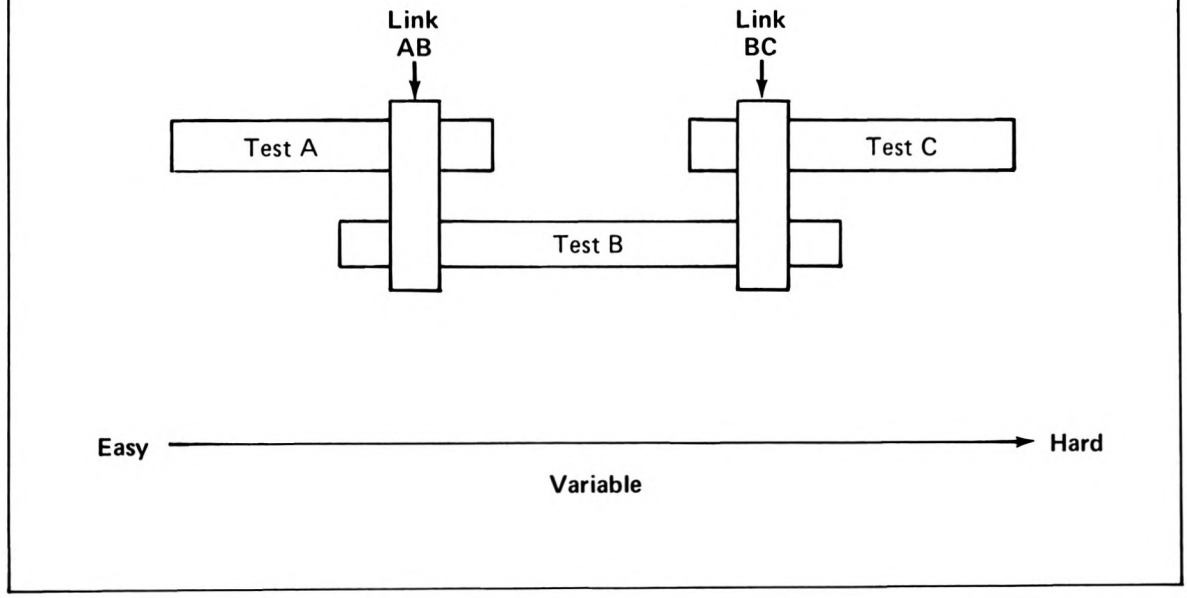
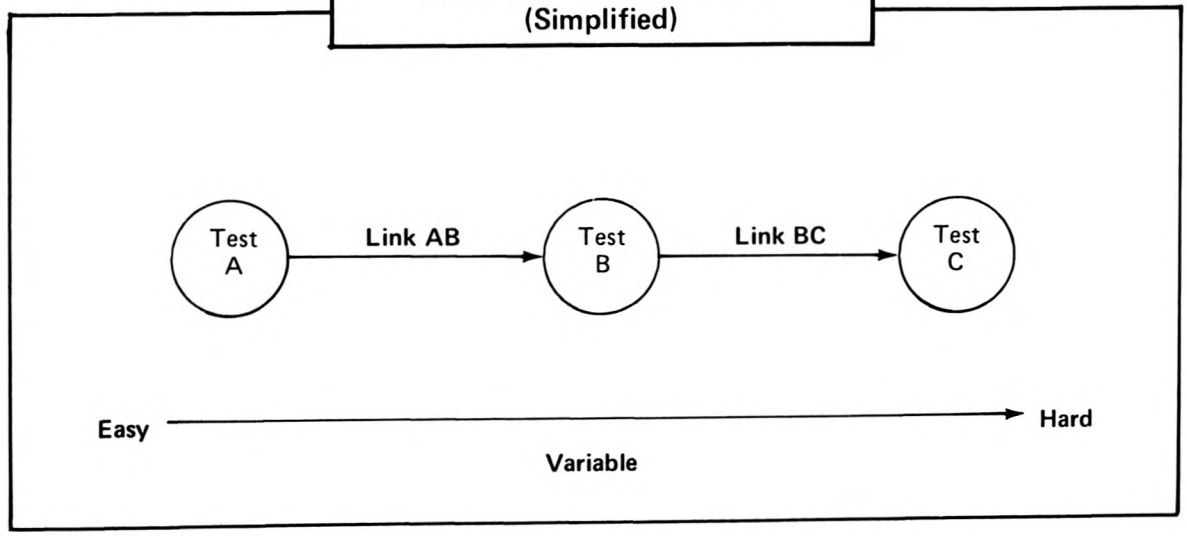


FIGURE 5.6.3

A CHAIN OF TWO LINKS (Simplified)



Three links can be constructed to form a loop as in Figure 5.6.4. This loop is an important linking structure because it yields an additional test of link coherence. If the three links in a loop are consistent, then the sum of their three link translations should estimate zero.

$$G_{AB} + G_{BC} + G_{CA} \simeq 0$$

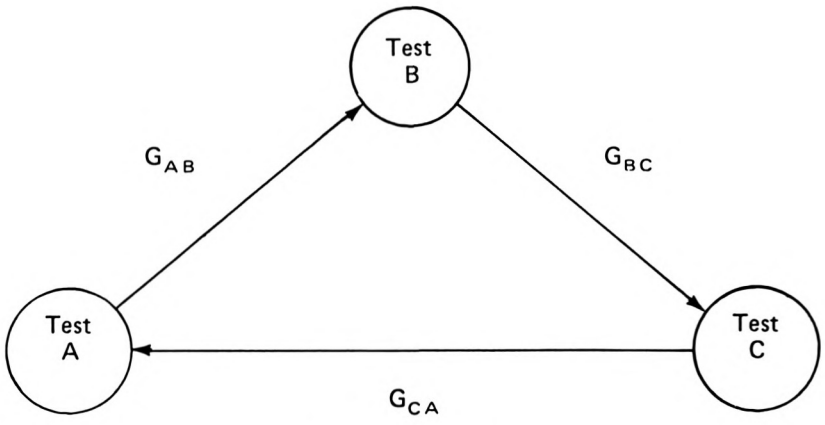
Notice that G_{AB} means the shift from Test A to Test B as we go around the loop clockwise so that G_{CA} means the shift from Test C back to Test A. Estimating zero "statistically" means that the sum of these shifts should come to within a standard error or two of zero. The standard error of the sum $G_{AB} + G_{BC} + G_{CA}$ will be about

$$3.5(1/N_{AB}K_{AB} + 1/N_{BC}K_{BC} + 1/N_{CA}K_{CA})^{1/2}$$

in which the N's are the calibration sample sizes and the K's are the number of items in each link.

FIGURE 5.6.4

A LOOP OF THREE LINKS



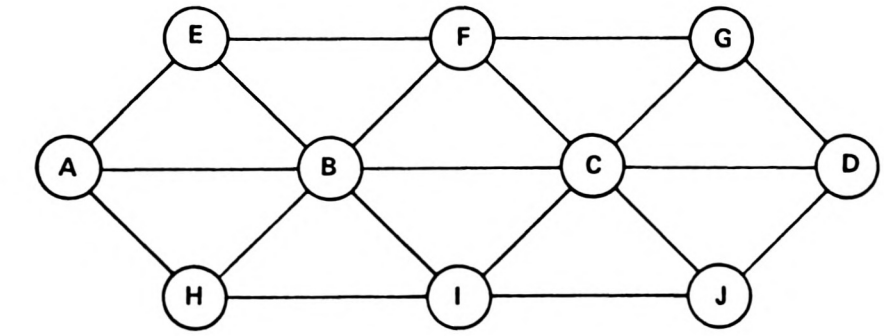
$$G_{AB} + G_{BC} + G_{CA} \simeq 0$$

$$SE(G_{AB} + G_{BC} + G_{CA}) \simeq 3.5\,(1/N_{AB}K_{AB} + 1/N_{BC}K_{BC} + 1/N_{CA}K_{CA})^{1/2}$$

With four or more tests we can construct a network of loops. For example, a sequence of increasingly difficult tests could be commonly calibrated by a series of connecting links as shown in Figure 5.6.5. These ten tests mark out seven levels of difficulty from Tests A through D. This network could connect ten 60-item tests by means of nineteen 10-item links to cover 600 - 190 = 410 items. If 200 persons were used for each test, then 410 items could be evaluated for possible calibration together from the responses of only 2,000 persons. Even 1,000 persons, at 100 per test, would provide a substantial purchase on the possibilities for building an item bank out of the best of the 410 items.

FIGURE 5.6.5

A NETWORK CONNECTING
TEN TESTS WITH NINETEEN LINKS



Easy → Hard

Variable

The building blocks of a test network are the loops of three tests each. If a loop fits the Rasch model, then its three translations should sum to within a standard error or two of zero. Thus the success of the network at linking item calibrations can be evaluated from the magnitudes and directions of these loop sums. Shaky regions can be identified and steps taken to avoid or improve them.

The implementation of test networks can lead to banks of commonly calibrated items far larger in number and far more dispersed in difficulty than any single person can handle. The resulting banks, because of the calibration of their items onto one common variable, can provide the item resources for a prolific family of useful tests, long or short, easy or hard, widely spaced in item difficulty or narrowly focused, all automatically equated in the measures they imply.

These methods for building item banks can be applied to existing tests, if they have been carefully constructed. Suppose we have two non-overlapping, sequential series of tests A_1, A_2, A_3, A_4 and B_1, B_2, B_3, B_4 which we want to equate by Rasch methods. All eight tests can be equated by connecting them with a new series of intermediate tests X, Y and Z made up entirely from items common to both series as shown in Figure 5.6.6. Were the A and B series of tests in Figure 5.6.6 still in the planning stage, they could also be linked directly by embedding common items in each test according to the pattern shown in Figure 5.6.7.

Since coherence is a vital concern in the building of an item bank, we are especially interested in linking structures which maximize statistical control over the joint coherence of all item calibrations. Networks which maximize the number of links among test forms so that each form is linked to as many other forms as possible do this. In the extreme, this leads to a web in which every individual item in a form links that form to another different form.

FIGURE 5.6.6

CONNECTING TWO NON-OVERLAPPING TEST SERIES BY INTERMEDIATE LINKING TESTS

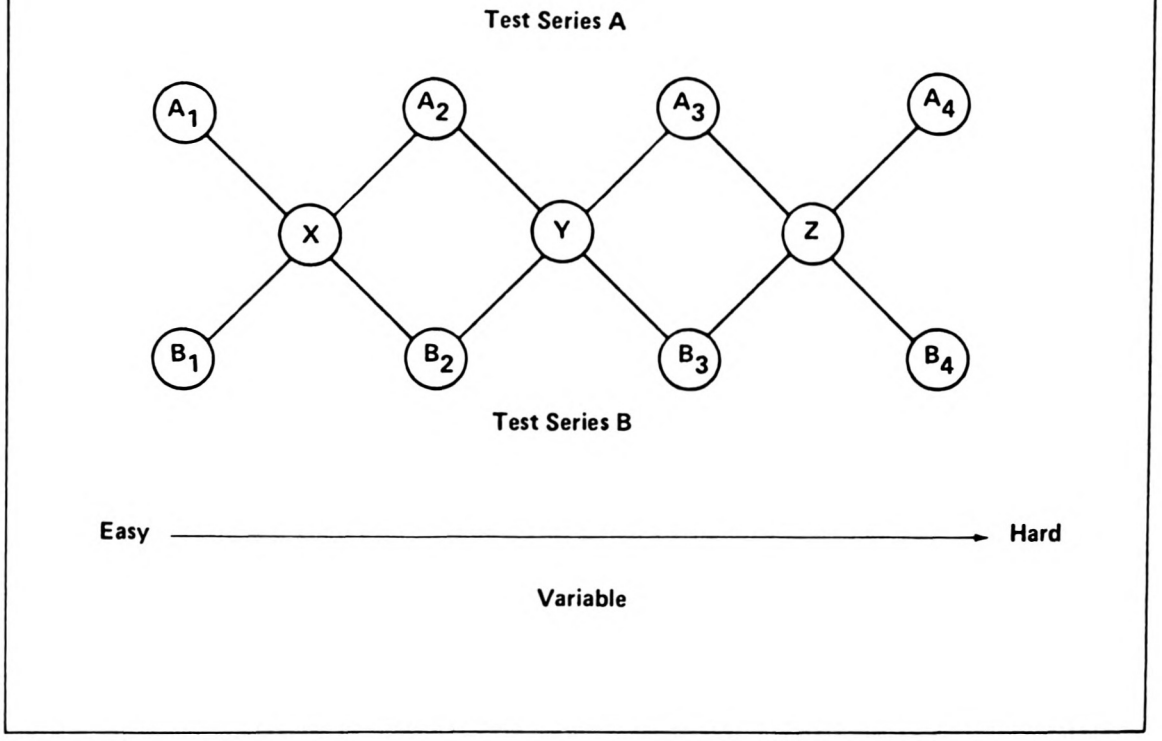
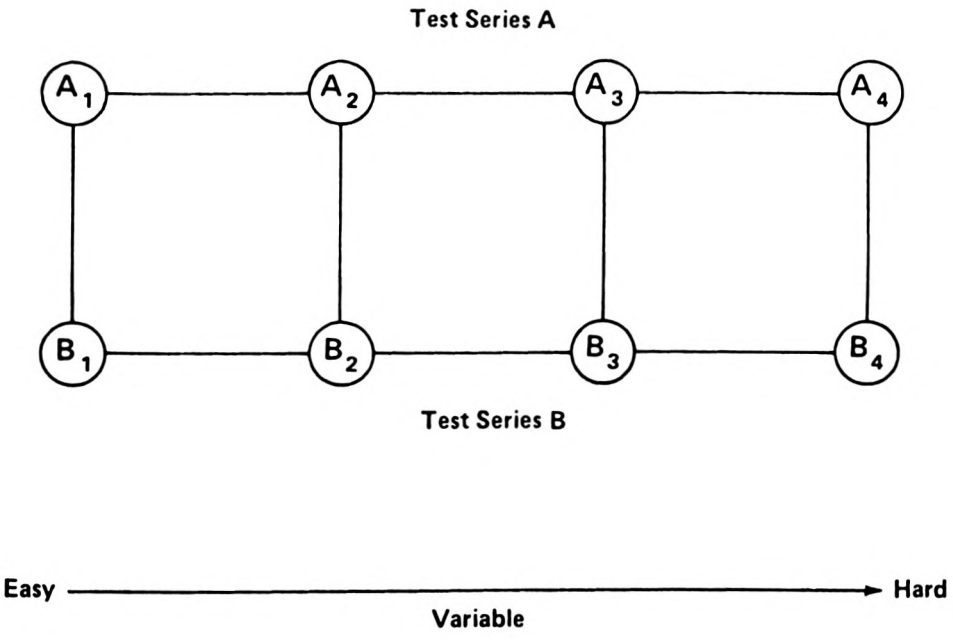


FIGURE 5.6.7

CONNECTING TWO TEST SERIES BY EMBEDDING COMMON ITEMS

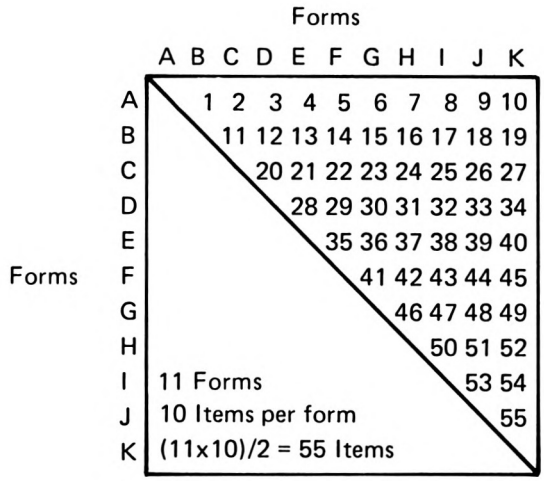


To illustrate we take a very small banking problem where we use 10 items per form in a web in which each of these 10 items also appears in one of 10 other different forms. The complete set of 10 + 1 = 11 forms constitutes a web woven out of 11 x 10/2 = 55 individual linking items. Every one of the 11 forms is woven to every other form. The pattern looks like the picture in Figure 5.6.8.

We will call this bank building design a "complete" web because every form is woven to every other form. In the design of useful webs, however, there are three constraints which affect their construction. These are the total number of items we want to calibrate into the bank, the maximum number of items which we can combine into a single form and the extent to which the bank we have in mind reaches out in difficulty beyond the capacity of any one person.

The testing situation and the capacity of the persons taking the test forms will limit the number of items we can put into a single form. It will usually happen, however, that we want to calibrate many more items than we can use up in a complete web like the one illustrated in Figure 5.6.8. There are two possibilities for including extra items. The simplest, but not the best statistically, is to design a "nuclear" complete web which uses up some portion of the items we can include in a single form. We then fill out the required form length with additional "tag" items. These tag items are calibrated into the bank along with the link items in their form. Unlike the link items, however, which always

FIGURE 5.6.8
A COMPLETE WEB FOR PARALLEL FORMS



The number entered in each cell is the identification of the item linking the two forms which define the position of that cell.

appear in two forms, the tag items appear in only one form and so give no help with linking forms together into one commonly calibrated bank.

Another possibility, which is better statistically, is to increase the number of forms used while keeping the items per form fixed at the required limit. This opens the web in a systematic way but still uses every item twice so that the paired data on that item can be used to evaluate the coherence of bank calibrations. Figure 5.6.9 shows an "incomplete" web for a 21 form design with 10 items per form, as in Figure 5.6.8, but with nearly twice as many items used in the incomplete web.

The incomplete web in Figure 5.6.9 is suitable for linking a set of parallel test forms. When the reach of the bank goes beyond the capacity of any one person, however, neither of the webs in Figures 5.6.8 and 5.6.9 will suffice, because we will be unable to combine items from the easy and hard ends of the bank into the same forms. The triangle of linking items in the upper right corners of Figures 5.6.8 and 5.6.9 will not be functional and will have to be deleted. In order to maintain the balance of linking along the variable we will have to do something at each end of the web to fill out the easiest and hardest forms so that the extremes are as tightly linked as the center. Figure 5.6.10 shows how this can be done systematically for a set of 21 sequential forms. We still have 10 items per form but now only adjacent forms are linked together. There are no common items connecting the easiest forms directly with the hardest forms. But over the range of the variable the forms near to one another in difficulty level are woven together with the maximum number of item links.

Each linking item in the webs shown in Figures 5.6.8, 5.6.9 and 5.6.10 could in fact refer to a cluster of two or more items which appear together in each of the two forms they link. Sometimes the design or printing format of items forces them into clusters. This happens typically in reading comprehension tests where clusters of items are attached to reading passages. It also occurs naturally on math and information retrieval tests where clusters of items refer to common graphs. Clustering, of course, increases the item length of each form by a factor equal to the cluster size.

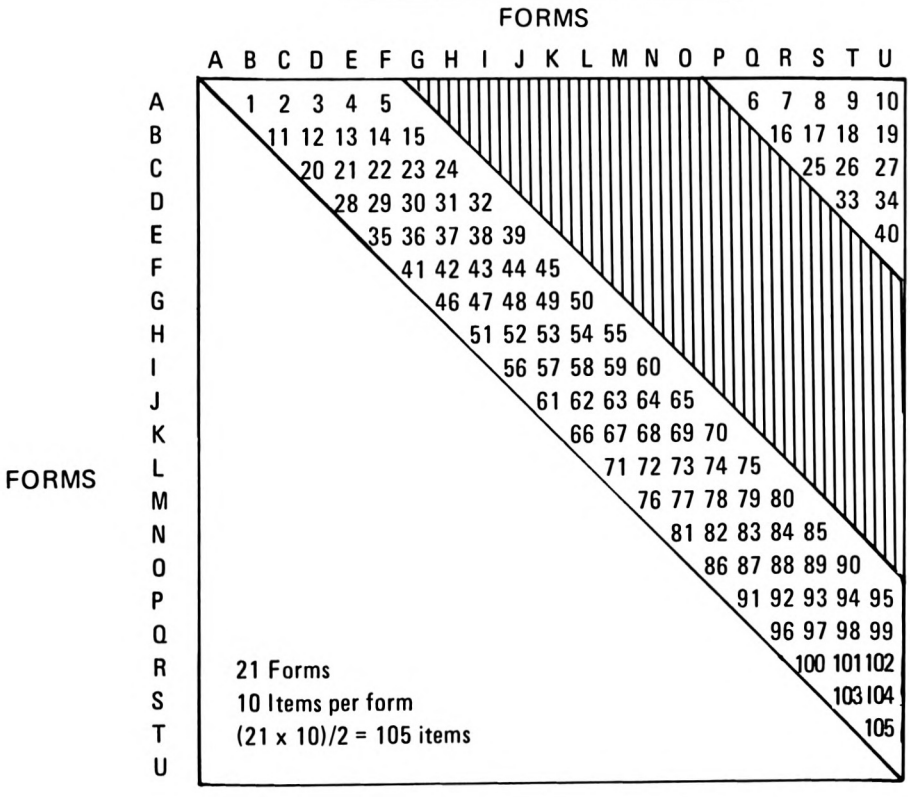
The statistical analysis of a bank-building web is simple, if the web is complete as in Figure 5.6.8. The row means of the corresponding matrix of form links are least square estimates of the form difficulties. We need only be careful about signs. If the web cell entry G_{jk} estimates the difference in difficulty $(\delta_j - \delta_k)$ between forms j and k and the form difficulties are centered at zero so that $\delta_\cdot = 0$, then

$$G_{j\cdot} = \sum_{k}^{M} G_{jk}/M \approx \delta_j$$

The row means of the link matrix calibrate the forms onto one common variable. Once form difficulties are obtained they need only be added to the item difficulties within forms to bring all items onto the common variable shared by the forms.

FIGURE 5.6.9

AN INCOMPLETE WEB FOR PARALLEL FORMS

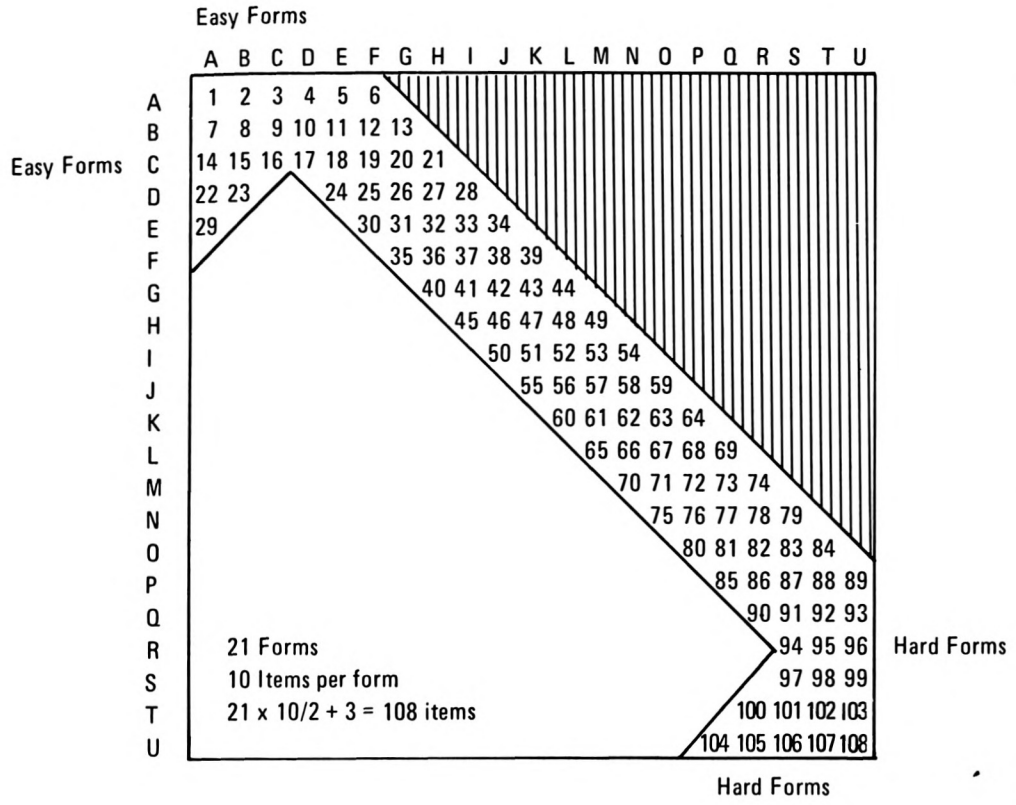


Formulation: $N = ML/2$

where
- N = number of items (or links) in the bank
- M = number of forms i.e., $2N/L$
- L = number of items (or links) per form must be even

FIGURE 5.6.10

AN INCOMPLETE WEB FOR SEQUENTIAL FORMS



Formulation: $N = ML/2 + K$

where
- N = number of items (or links) in the bank
- M = number of forms i.e., $2(N - K)/L$
- L = number of items (or links) per form must be even
- K = $L/4$, if $L/2$ is even
  = $(L + 2)/4$, if $L/2$ is odd

The incomplete webs in Figures 5.6.9 and 5.6.10 require us to estimate row means from a matrix with missing data. The skew symmetry of link matrices helps the solution to this problem which can be done satisfactorily by iteration or regression.

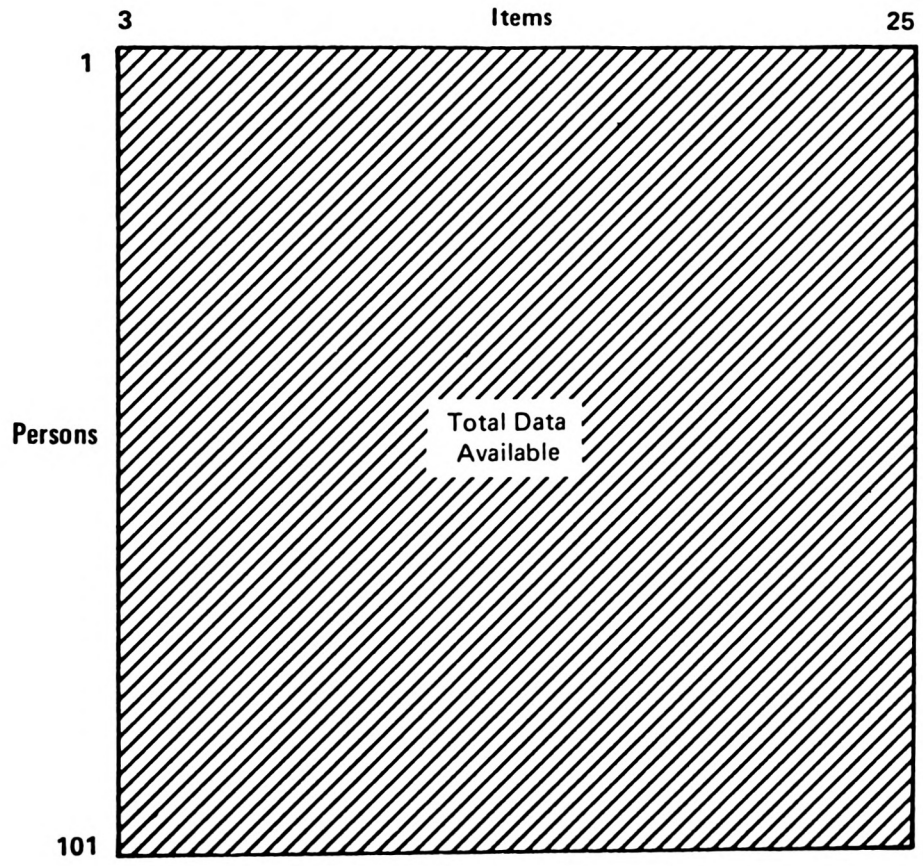
5.7 BANKING THE KCTB DATA

The KCTB is a short test so it was practical to ask all 101 persons to attempt all 23 items giving us the response matrix illustrated in Figure 5.7.1. However, most item banking projects involve the calibration of hundreds of items given to thousands of examinees. It is then impossible to ask every person to take every item. Fortunately building an item bank does not require such an undertaking. As we saw in Section 5.6, items can be joined together by a network of links. In general, two types of form equating are possible, common persons and common items.

One way to link separate forms is to administer them both to the same sample of persons. We illustrate "common person" equating with our KCTB data by defining two non-overlapping sequential tests, EASY and HARD, and finding everyone who produced measurable responses simultaneously in both tests. This is an attempt at the vertical equating of an easy and a hard test and we can expect persons with usable scores on both tests to be scarce. With our KCTB example there are only 29 such persons out of 101. The picture of this common person equating in Figure 5.7.2 shows the core of 29 persons from the total sample linking two non-overlapping parts of the KCTB, a 9-item EASY test and an 8-item HARD test.

FIGURE 5.7.1

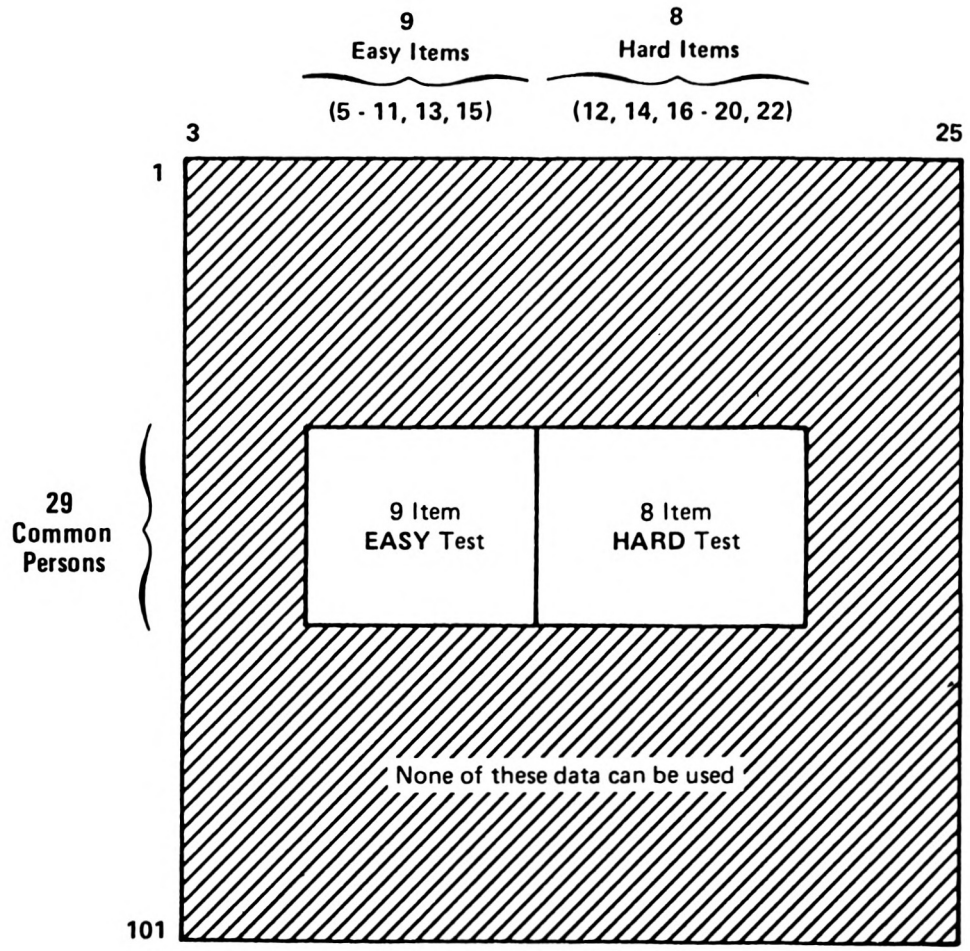
COMMON PERSONS AND COMMON ITEMS WITH KCTB



Item 2 has been dropped because it is too easy to be useful.

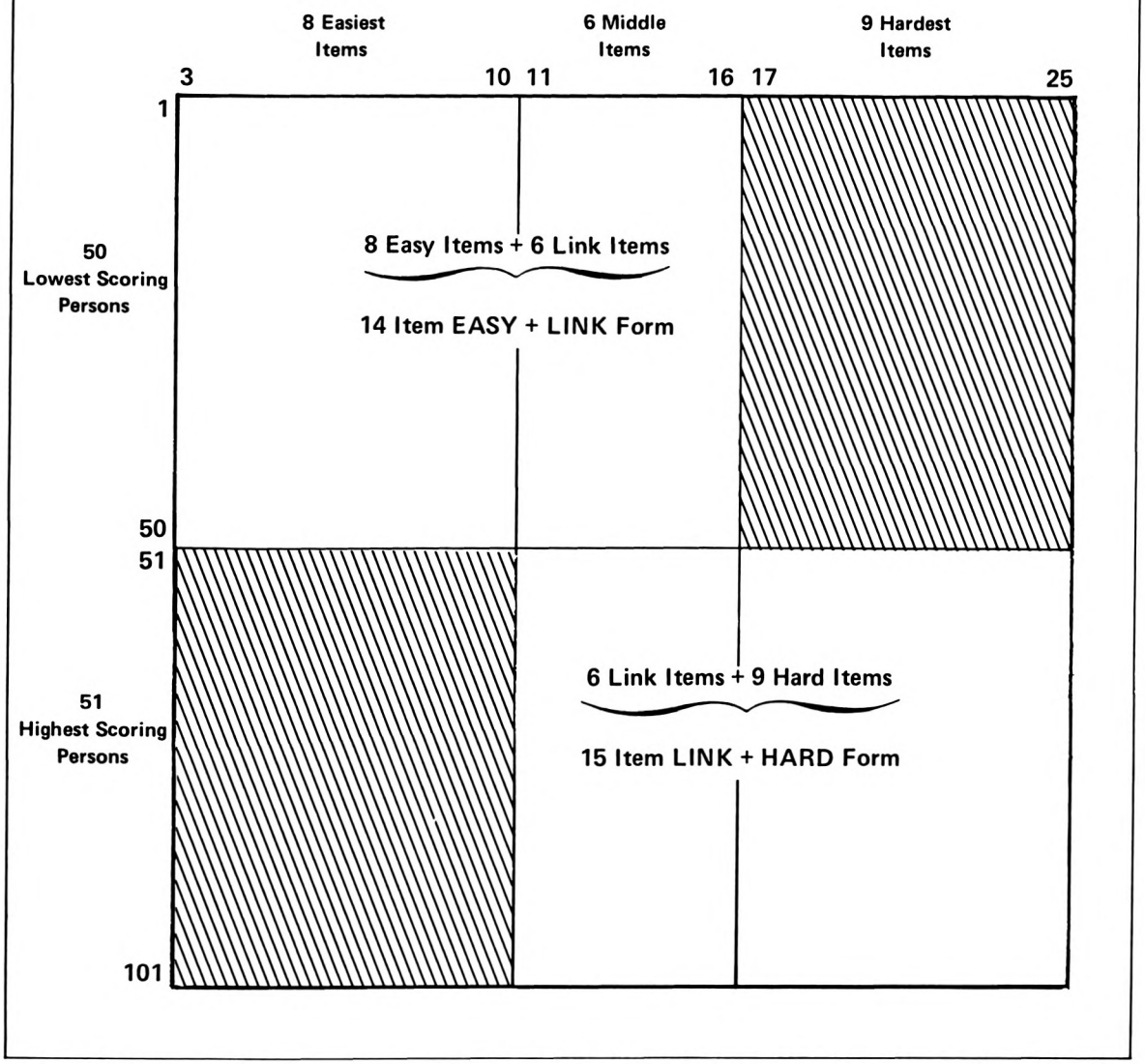
FIGURE 5.7.2

COMMON PERSON EQUATING
WITH KCTB



A better way to equate forms is by using common items. This approach to KCTB is shown in Figure 5.7.3. There we show eight easy items connected to nine hard items by a six item link producing a 14 item EASY + LINK form taken by the 50 lowest scoring persons and a 15 item LINK + HARD form taken by the 51 highest scoring persons.

FIGURE 5.7.3
COMMON ITEM EQUATING WITH KCTB



5.8 COMMON PERSON EQUATING WITH THE KCTB

Among the 101 persons taking KCTB Items 3 through 25 we found two non-overlapping sequential forms, called EASY and HARD, for which 29 persons had a pair of usable scores. The EASY form was made from Items 5 through 11, 13 and 15. The HARD form was made from Items 12, 14, 16 through 20 and 22. The 29 persons were those who remained after high scoring persons were removed because of perfect scores on the EASY test and low scoring persons were removed because of zero scores on the HARD TEST.

The measurements of these 29 persons on each form constitute the common person data for linking the EASY and HARD forms together. It is the difference in the two ability means which estimates the shift required to bring the EASY and HARD forms onto a common scale. The ability statistics for the 29 persons on each form are

	EASY Form	HARD Form	Difference
Mean Ability	1.49	-0.57	2.06
Standard Deviation	0.80	0.43	

The equating procedure is as follows:

1. Use the observed difference in sample mean ability 1.49 - (-0.57) = 2.06 as the estimated difficulty difference between the two forms.
2. Apportion this difference over the nine EASY items and the eight HARD items so that the average difficulty of all 17 items becomes zero.

   For the nine EASY items use

   $$[(17 - 9)/17] (2.06) = 0.97$$

   For the eight HARD items use

   $$[(17 - 8)/17] (2.06) = 1.09$$

3. Bring the two forms onto a common scale by subtracting 0.97 from each EASY form item difficulty and adding 1.09 to each HARD form item difficulty.

These computations are displayed in Table 5.8.1. Column 1 gives the KCTB item name for the 17 items used in the EASY and HARD forms. Column 2 gives the separate item calibrations for the EASY form. Column 3 gives the separate calibrations for the HARD form. Because these separate calibrations are each centered within their own form Columns 2 and 3 each sum to zero.

Converting the calibrations in Columns 2 and 3 to a centered common person scale requires subtracting 0.97 from the EASY form item difficulties in Column 2 and adding 1.09 to the HARD form item difficulties in Column 3. This is done in Columns 4 and 5 resulting in a common person scale for all 17 items centered at 0.0.

In order to evaluate the efficacy of this common person equating we obtained a combined calibration of all 17 items from the same 29 persons. Column 6 gives these reference calibrations and Column 7 gives the differences between the common person scale and the reference scale.

Figure 5.8.1 compares the common person scale and the reference scale. The small differences between the two scales show that the common person technique can produce results equivalent to a combined calibration of both tests.

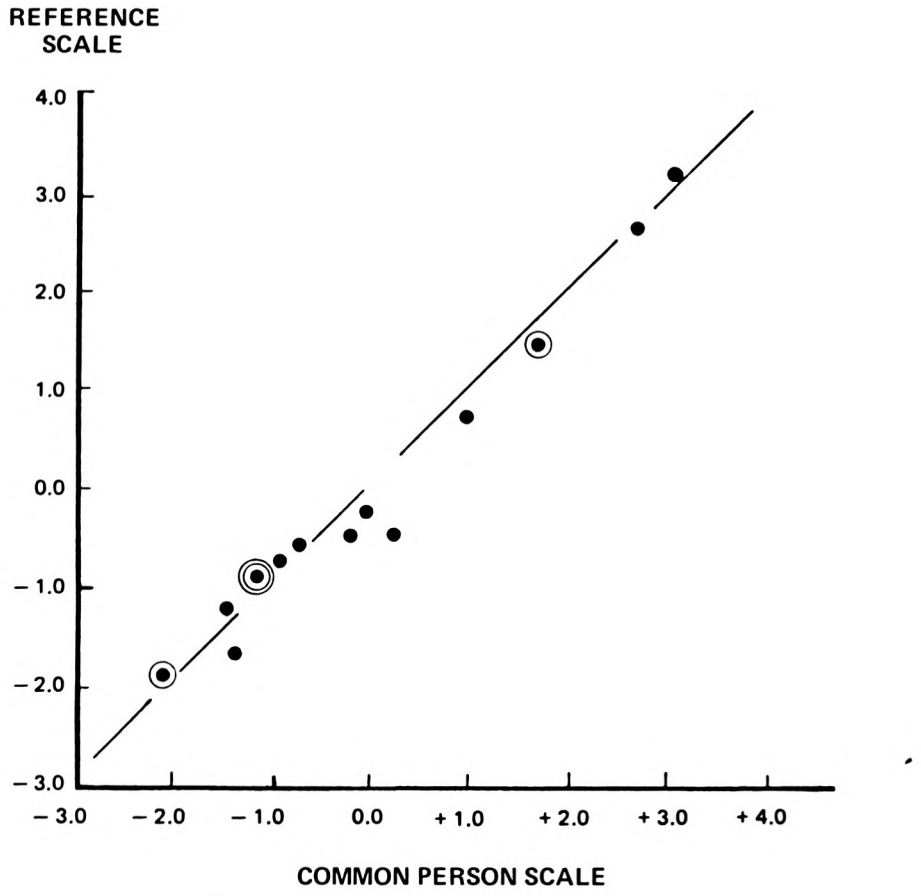
TABLE 5.8.1
EQUATING EASY AND HARD FORMS USING COMMON PERSONS

1	2	3	4	5	6	7
	Separate Calibrations		Common Person Scale		Reference	
Item Name	EASY d_E	HARD d_H	EASY $d_C=d_E-0.97$	HARD $d_C=d_H+1.09$	Calibration* d_R	Difference d_C-d_R
5	0.03		−0.94		−1.04	−0.10
6	0.03		−0.94		−1.04	−0.10
7	−0.94		−1.91		−2.05	−0.14
8	0.03		−0.94		−1.04	−0.10
9	0.24		−0.73		−0.82	−0.09
10	0.43		−0.54		−0.62	−0.08
11	1.36		0.39		0.35	−0.04
12		−1.44		−0.32	−0.10	0.22
13	−0.22		−1.19		−1.30	−0.11
14		−1.25		−0.16	0.05	0.21
15	−0.94		−1.91		−2.05	−0.14
16		−2.66		−1.57	−1.30	0.27
17		−0.12		0.97	1.10	0.13
18		0.65		1.74	1.81	0.07
19		0.65		1.74	1.81	0.07
20		1.83		2.92	2.90	−0.02
22		2.32		3.41	3.36	−0.05
Mean	0.00	0.00	0.00		0.00	0.00
Standard Deviation	0.70	1.70	1.62		1.65	0.14

* Based on 29 persons taking all 17 items

FIGURE 5.8.1

COMPARISON OF COMMON PERSON SCALE WITH THE REFERENCE SCALE



5.9 COMMON ITEM EQUATING WITH THE KCTB

To illustrate common item equating we have divided the 23 KCTB items into three parts: EASY, LINK and HARD. The EASY + LINK form contains eight EASY items and six LINK items to make a 14 item easy test. The LINK + HARD form contains the six common LINK items plus nine HARD items making a 15 item hard test.

Each of these forms was calibrated on separate samples. The EASY + LINK form was calibrated on the 50 lowest scoring persons and the LINK + HARD form was calibrated on the 51 highest scoring persons. These calibrations are given in Table 5.9.1.

The paired calibrations of the six linking items, 11 through 16, are given again in Columns 2 and 3 of Table 5.9.2. Their differences $D = d_E - d_H$ are given in Column 4. The mean of these differences is 4.11 which is the difficulty difference between the EASY + LINK form and the LINK + HARD form. When this difference of 4.11 is subtracted from D we have the residuals from linking given in Column 5.

TABLE 5.9.1
ITEM CALIBRATIONS OF EASY + LINK AND LINK + HARD FORMS

Item Name	EASY + LINK Difficulty	EASY + LINK Error	LINK + HARD Difficulty	LINK + HARD Error
3	- 3.80			
4	- 2.00			
5	- 0.37			
6	- 0.37			
7	- 2.00			
8	- 0.37			
9	0.06			
10	0.20			
11	0.97	.36	- 2.24	.49
12	2.08	.38	- 1.83	.44
13	1.58	.36	- 3.22	.73
14	1.95	.37	- 2.80	.61
15	0.84	.36	- 3.90	1.01
16	1.21	.36	- 2.02	.46
17			0.60	
18			- 0.50	
19			0.26	
20			1.18	
21			1.56	
22			1.56	
23			2.78	
24			4.51	
25			4.06	
Mean	0.00		0.00	
Standard Deviation	1.68		2.64	

If these items are providing a usable link, their residuals should distribute around zero with the standard error predicted by the model. The standard errors

$$S_D = (S_E^2 + S_H^2)^{1/2}$$

of these residuals are given in Column 6 and the standardized residuals

$$z = (D - 4.11)/S_D$$

are given in Column 7.

Figures 5.9.1 is a plot of the EASY calibrations of these LINK items against their HARD calibrations. The item points are well within 95% control lines demonstrating that the shift estimated from this link can be used to connect the two forms.

TABLE 5.9.2
LINK ANALYSIS

1	2	3	4	5	6	7
	Calculating LINK SHIFT			Testing LINK FIT		
Item Name	EASY d_E	HARD d_H	Difference $D = d_E - d_H$	Residual Difference $D - 4.11$	Standard Error of Residual S_D	Standardized Residual $z = (D - 4.11)/S_D$
11	0.97	- 2.24	3.21	- 0.90	0.61	- 1.48
12	2.08	- 1.83	3.91	- 0.20	0.58	- 0.34
13	1.58	- 3.22	4.80	0.69	0.81	0.85
14	1.95	- 2.80	4.75	0.64	0.71	0.90
15	0.84	- 3.90	4.74	0.63	1.07	0.59
16	1.21	- 2.02	3.23	- 0.88	0.58	- 1.52
Mean	1.44	- 2.67	4.11	0.00		$-0.17 \simeq 0$
Standard Deviation	0.52	0.79	0.76	0.76		$1.13 \simeq 1$

$$\text{LINK Shift} = \sum_{i}^{6} D_i/6 = 4.11$$

Standard Error of Residual: $S_D = (S_E^2 + S_H^2)^{1/2}$

Expected mean of z is 0

Expected standard deviation of z is 1

Our next step is to connect EASY + LINK to LINK + HARD. We do this by connecting both LINKs and HARD to EASY. Table 5.9.3 shows the method used. In Column 1 we have the item name for each of the 23 KCTB items. The item difficulties of Items 3 through 10 are given in Column 2. Because we will reference all other items to EASY, we record the difficulties for Item 3 through 10 directly into Column 6. For LINK Items 11 through 16 we have two sets of difficulties. In Column 2 we have difficulty estimates for Items 11 through 16 from calibration with the EASY items. In Column 3 we have difficulty estimates for these same items obtained from their calibration with the HARD items.

To the LINK difficulties d_H we add the link difficulty difference of 4.11. Then we average the LINK d_E difficulties with the LINK d_H difficulties that were adjusted by the LINK shift of 4.11. The average of the two LINK estimates $(d_E + d_H + 4.11)/2$ for Items 11 through 16 is given in Column 5. We enter these in Column 6.

FIGURE 5.9.1

LINK FOR COMMON ITEM EQUATING

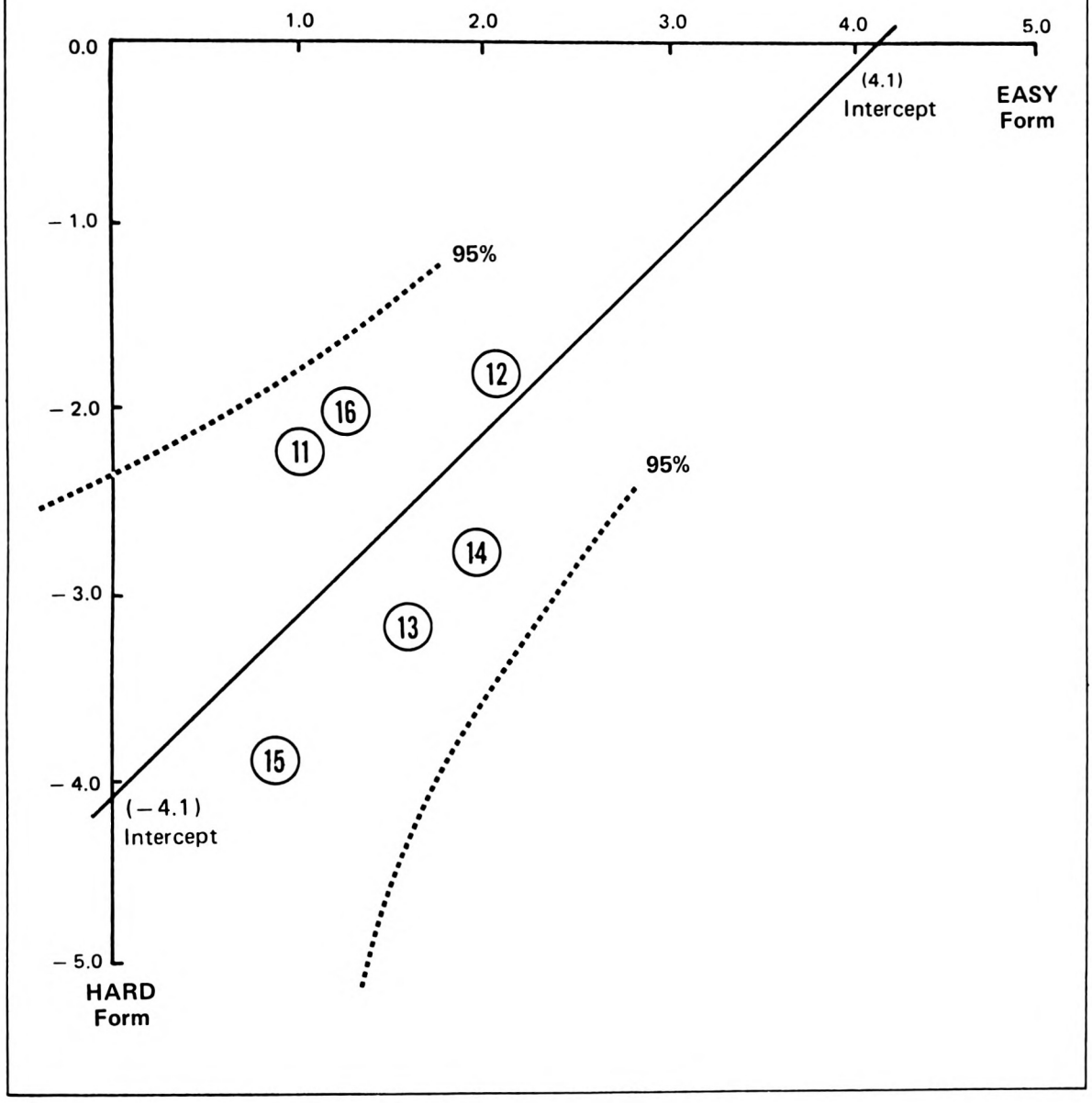


TABLE 5.9.3.
EQUATING EASY AND HARD FORMS BY A COMMON ITEM LINK

1	2	3	4	5	6	7
	Calibrating Each Form		Shifting to EASY + LINK		Common Item Scale	
Item Name	EASY + LINK d_E	LINK + HARD d_H	$d_H + 4.11$	$(d_E + d_H + 4.11)/2$	d_C	Centered $d_C - 2.30$
3	-3.80				- 3.80	- 6.10
4	-2.00				- 2.00	- 4.30
5	-0.37				- 0.37	- 2.67
6	-0.37				- 0.37	- 2.67
7	-2.00				- 2.00	- 4.30
8	-0.37				- 0.37	- 2.67
9	0.06				0.06	- 2.20
10	0.20				0.20	- 2.10
11	0.97	-2.24	1.87	1.42	1.42	- 0.92
12	2.08	-1.83	2.28	2.18	2.18	- 0.12
13	1.58	-3.22	0.89	1.24	1.24	- 1.06
14	1.95	-2.80	1.31	1.63	1.63	- 0.67
15	0.84	-3.90	0.21	0.53	0.53	- 1.77
16	1.21	-2.02	2.09	1.65	1.65	- 0.65
17		0.60	4.71		4.71	2.41
18		-0.50	3.61		3.61	1.31
19		0.26	4.37		4.37	2.07
20		1.18	5.29		5.29	2.99
21		1.56	5.67		5.67	3.37
22		1.56	5.67		5.67	3.37
23		2.78	6.89		6.89	4.59
24		4.51	8.62		8.62	6.32
25		4.06	8.17		8.17	5.87
Mean	0.00	0.00	4.11		2.30	0.00
Standard Deviation	1.68	2.64	2.64		3.37	3.37

Finally in order to place the HARD items on the common scale we add 4.11 to HARD Items 17 through 25 and bring these difficulty estimates over to complete Column 6. We then have in Column 6 a new common item scale with the average of two LINK difficulty estimates and the HARD difficulty estimates all connected to the EASY item difficulty estimates.

The mean of this common item scale in Column 6 is 2.30 so we subtract 2.30 from each item difficulty in Column 6 to center the new scale at 0.00 as shown in Column 7.

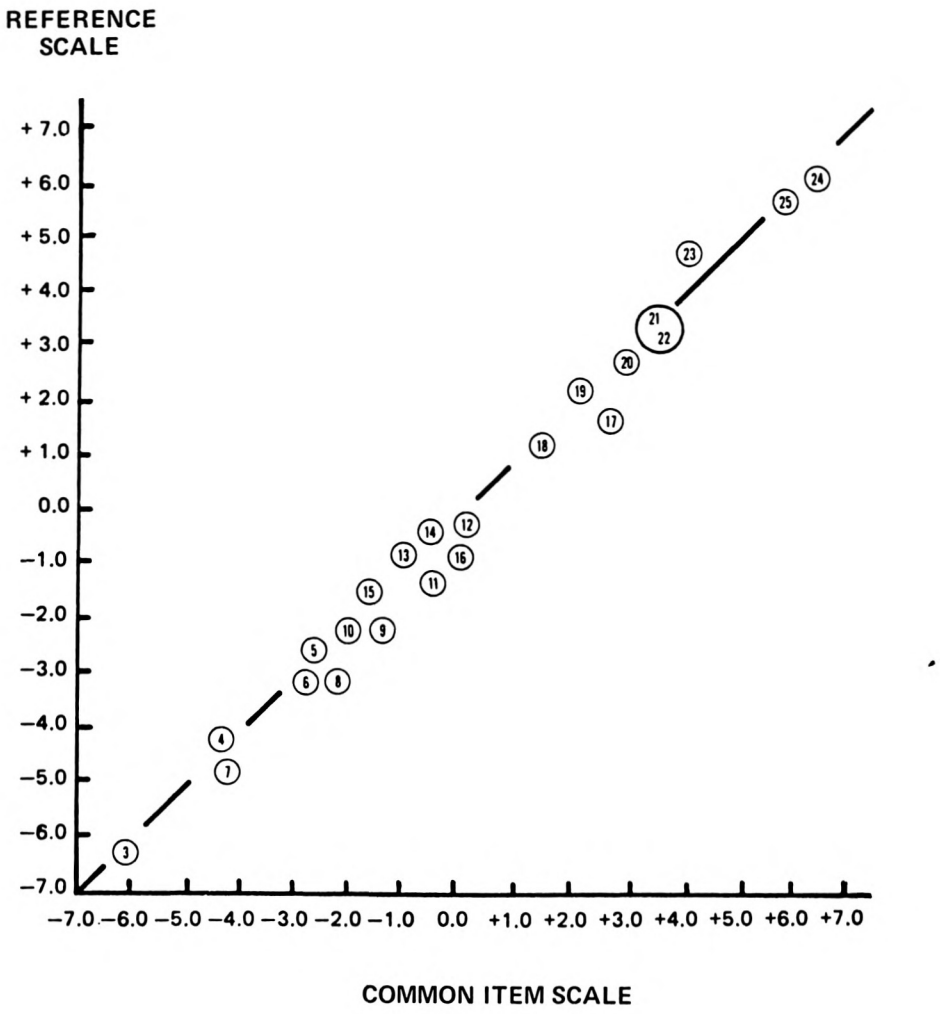
To assess the adequacy of this common item equating we will compare it to the item difficulties we would have gotten had we not attempted linking but used all 101 person responses to all 23 items. The common item difficulties from Table 5.9.3 are given in Column 2 of Table 5.9.4. Column 3 gives the reference calibrations of all 23 items from all 101 persons, and Column 4 shows the differences between the common item difficulties d_C and the reference scale item difficulties d_R. The plot of these values given in Figure 5.9.2 shows the items close to the expected identity line.

TABLE 5.9.4
COMPARING COMMON ITEM EQUATING WITH THE REFERENCE SCALE

1 Item Name	2 Common Item Scale $d_C - 2.30$	3 Reference Scale d_R	4 Difference $d_C - d_R$
3	- 6.10	-6.20	.10
4	- 4.30	-4.11	- .19
5	- 2.67	-2.58	- .09
6	- 2.67	-2.72	.05
7	- 4.30	-4.34	.04
8	- 2.67	-2.58	- .09
9	- 2.24	-2.06	- .18
10	- 2.10	-2.06	- .04
11	- 0.92	-1.03	.11
12	- 0.12	-0.12	.00
13	- 1.06	-0.85	- .21
14	- 0.67	-0.52	- .15
15	- 1.77	-1.52	- .25
16	- 0.65	-0.77	.12
17	2.41	1.93	.48
18	1.31	1.36	- .05
19	2.07	2.01	.06
20	2.99	2.88	.11
21	3.37	3.33	.04
22	3.37	3.33	.04
23	4.59	4.52	.07
24	6.32	6.27	.05
25	5.87	5.81	.06
Mean	0.00	0.00	0.00
Standard Deviation	3.37	3.32	0.15

FIGURE 5.9.2

COMPARISON OF COMMON ITEM SCALE WITH THE REFERENCE SCALE



5.10 CRITERION REFERENCING THE KCT VARIABLE

By locating all 23 KCTB items on a single scale we can make the definition of the KCT variable more explicit. These items which now mark out the variable are constructed out of a few basic components: number of taps, number of reverses and overall distance across blocks. It is the way these underlying components evolve along the variable which documents for us what a measure on the KCT variable means. Figure 5.10.1 gives the difficulty level of the KCTB items together with their number of taps, reverses and distances.

FIGURE 5.10.1

DOCUMENTING THE KCT VARIABLE (101 Persons By 28 Items)

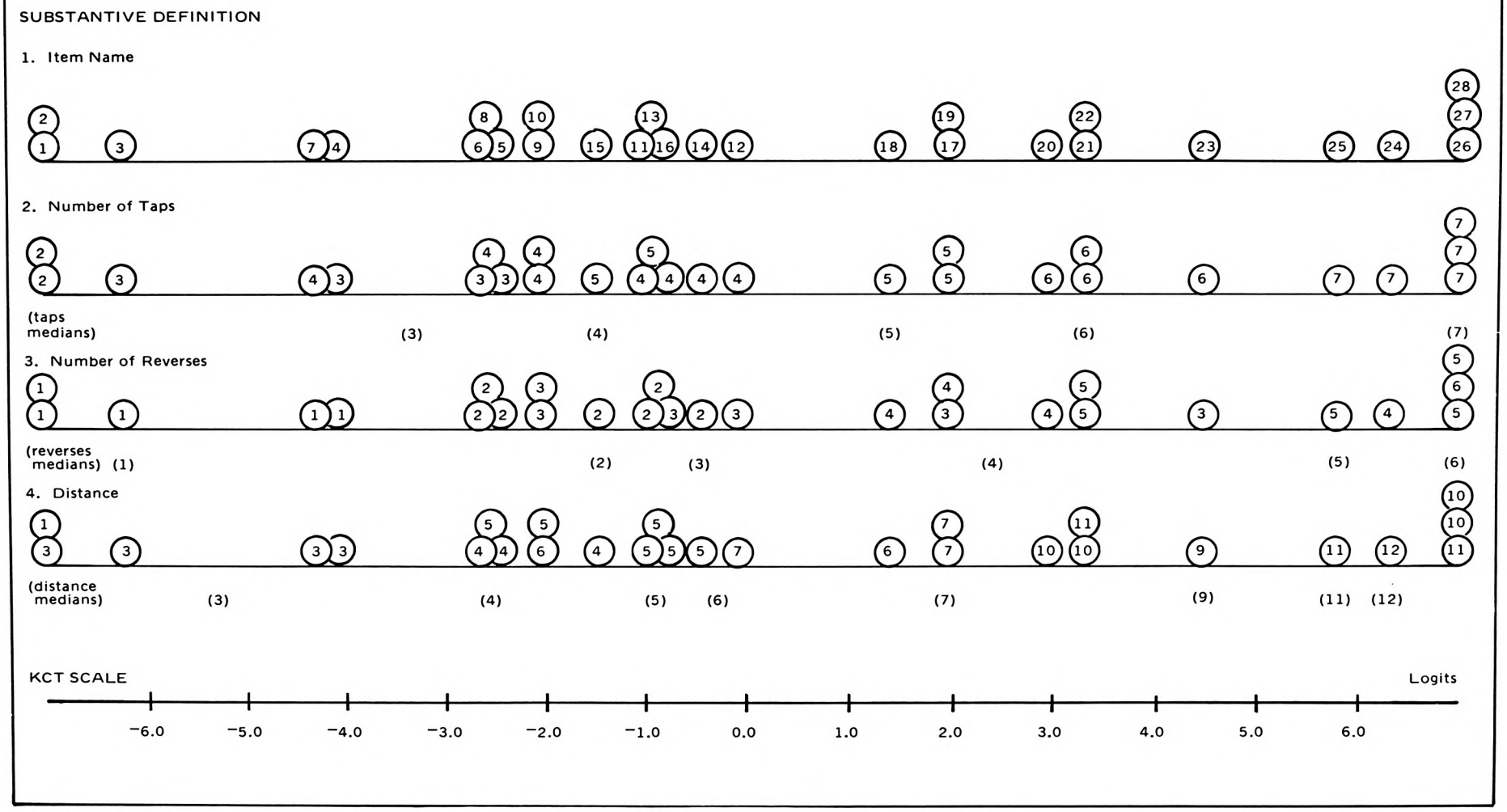
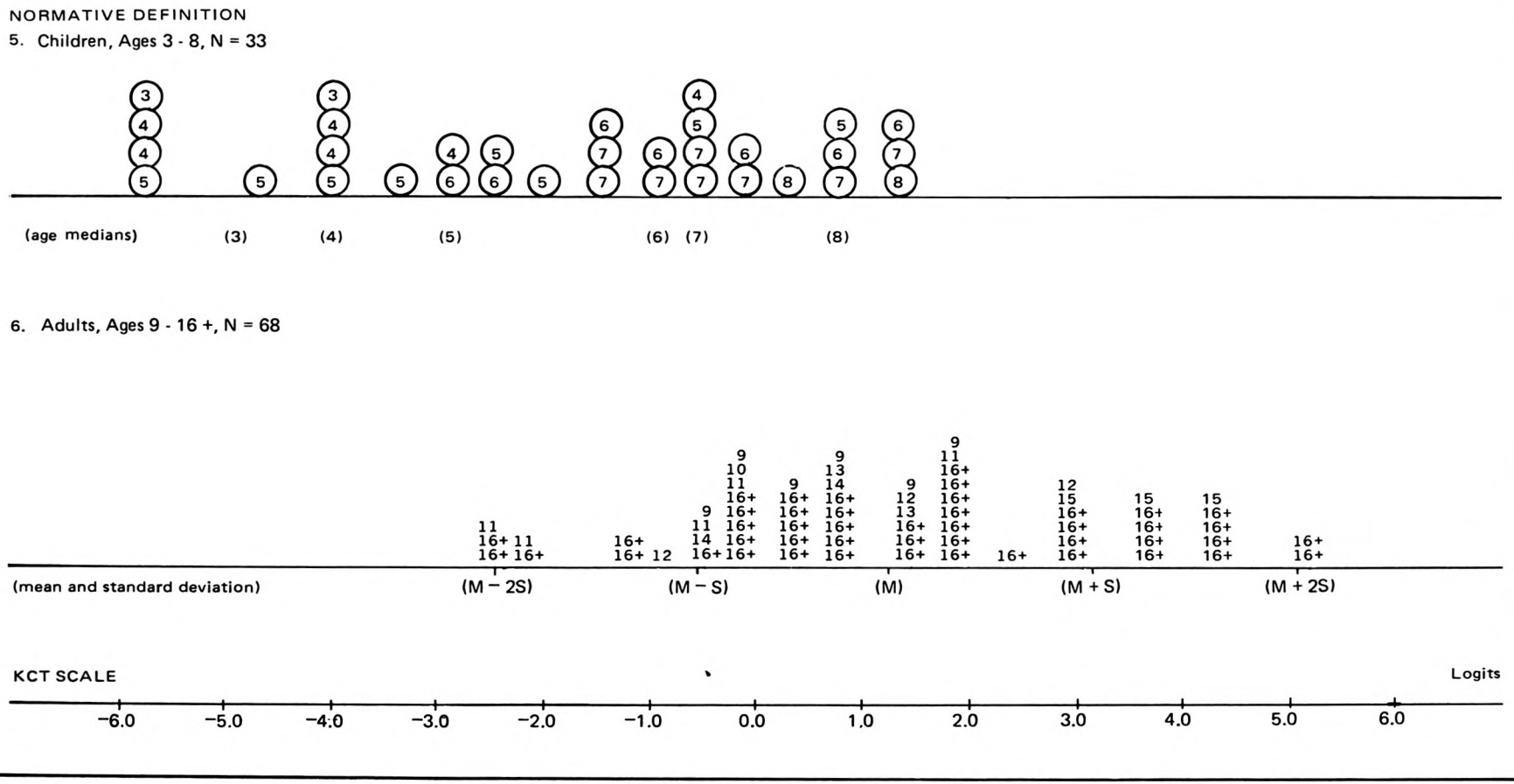


FIGURE 5.10.1

DOCUMENTING THE KCT VARIABLE (Continued)



Row 1 of Figure 5.10.1 contains the items of KCTB arranged by their calibrations on the variable according to the logit scale given at the bottom of the figure. The number of taps for each item is given in Row 2. Items 1 and 2 are two-tap items passed by all 101 persons in the sample. As we move up the variable, the number of taps goes from two to seven. Below Row 2 we have marked the median difficulty level for each number of taps from two to seven. Row 3 shows the number of reverses in each item and their median difficulty levels. Row 4 shows the distances in blocks tapped for each tapping series and their medians.

The pattern of taps, reverses and distances in Figure 5.10.1 show how the KCT variable is built out of these basic operations. This provides a substantive, or criterion, reference for the KCT variable. The resulting picture gives us insight into the nature of the variable which reaches beneath the individual items. In particular it shows us how to generate more items at any designated difficulty level.

We can also learn about the KCT variable by seeing how the 101 persons in our sample are distributed along it. In Rows 5 and 6 we show each person's position on the variable by their age in years. This allows us to norm reference the variable with age medians from three to eight years and to give an age distribution of "mature" persons of 9 or more years of age with a mean at 1.3 logits and a standard deviation of 1.9 logits. Thus Figure 5.10.1 becomes a map of the variable which is both criterion and norm referenced.

5.11 ITEM CALIBRATION QUALITY CONTROL

We cannot expect the items in a bank to retain their calibrations indefinitely or to work equally well for every person with whom they may be used. The quality of item calibration must be supervised continuously. This can be done conveniently by a routine examination of the differences between how persons actually respond to particular items and how we expect them to respond given our calibrations of the items and our measurements of the persons. These differences are residuals from expectation. An occasional surprising item residual suggests an anomalous testing situation or a peculiar person. Trends in item residuals, however, may be indicative of item failure. Tendencies for items to run into trouble, to shift difficulty or to be biased for some types of persons can be exposed by a cumulative analysis of item residuals over time, place and person type. Problematic items can then be removed from use or brought up-to-date in difficulty.

The purpose of item quality control is to maintain supervision over item calibration stability against the possible influences of age, sex, education or any other factor which might disturb item functioning. A quality control procedure requires that item usage be accompanied by concomitant educational and demographic information so as to provide a basis for analyzing whether these other variables threaten the stability of item calibration and hence disturb the interpretation of test responses. The discussion which follows builds on the analysis of fit developed in Chapter 4.

To implement item quality control we save from each use of an item:

x_{vi} the response 0 or 1 of person v to item i,

b_v the ability estimate of person v derived from their score on whatever "test" of calibrated items they took and

(y_v) the vector of demographic information which characterizes person v.

When the two pieces of information x_{vi} and b_v are combined with the item's bank difficulty d_i we can form a standardized residual z_{vi} which will retain all the information in this use of item i which bears on the possibility of a disturbance in its functioning.

In general this estimated residual z_{vi} is

$$z_{vi} = (x_{vi} - p_{vi})/[p_{vi}(1 - p_{vi})]^{1/2}$$

where

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)]$$

is the estimated probability of success for person v on item i and hence the estimated expected value of x_{vi} given the model.

Since x_{vi} can only take one of two values, 0 for an incorrect response or 1 for a correct response, the possibilities for z_{vi} and its square z_{vi}^2 are limited to those given in Table 5.11.1. The improbability of any particular response x_{vi}, as a function of its z_{vi}^2, is $1/(1 + z_{vi}^2)$. In the KCTB example there are 101 persons taking 23 items. These 23 x 101 item-by-person responses imply 2323 occasions for misfit. However, misfit can only show up when the difference between person ability b_v and item difficulty d_i is large enough so that one of the possible values for the response x_{vi} becomes significantly improbable. For this to happen the difference $(b_v - d_i)$ must be at least three logits. As a result there are only about 500 item-by-person occasions where misfit could occur.

TABLE 5.11.1

STANDARDIZED RESPONSE RESIDUALS

Response Value x_{vi}	Standardized Residuals: As a normal deviate $z_{vi} \sim N(0,1)$	Standardized Residuals: As a chi-square $z_{vi}^2 \sim \chi_1^2$
"Incorrect" 0	$z_{vi} = -p_{vi}/[p_{vi}(1 - p_{vi})]^{1/2}$ $= -[p_{vi}/(1 - p_{vi})]^{1/2}$ $= -\exp[(b_v - d_i)/2]$	$z_{vi}^2 = \exp(b_v - d_i)$
"Correct" 1	$z_{vi} = (1 - p_{vi})/[p_{vi}(1 - p_{vi})]^{1/2}$ $= [(1 - p_{vi})/p_{vi}]^{1/2}$ $= \exp[(d_i - b_v)/2]$	$z_{vi}^2 = \exp(d_i - b_v)$

Table 5.11.2 gives a summary of the unexpected responses observed in the KCTB data. Column 1 gives the range of absolute difference between person ability and item difficulty. Column 2 expresses this difference as $z^2 = \exp(|\, b - d \,|)$ and Column 3 converts z^2 to the response improbability $[1/(1 + z^2)]$ it implies.

TABLE 5.11.2
SUMMARY OF UNEXPECTED RESPONSES ON KCTB
101 PERSONS BY 23 ITEMS

Ability-Difficulty Difference	z^2	Improbability $1/(1 + z^2)$	Possible Count	Expected Count	Observed Count	Item Names	Person Names
Over 4.6	Over 99	Under .01	226	2	3	4, (7,) 9	(49M,) 68F, 93F
3.9 - 4.6	49 - 99	.02 - .01	133	5	3	8, 10, 12	79M 83F, [95M]
2.9 - 3.8	19 - 49	.05 - .02	184	20	8	[3,] [3,] 5	12M, 13M, 27F
						6, (7,) 11	47M, (49M,) 82F
						18, 19	[95M] 10F

We have counted the number of item-by-person interactions which could fall within each row of Table 5.11.2 and multiplied this "possible" count by its improbability to estimate the count we might expect if these data fit the model. This was done by multiplying (226) x .01 $\cong$ 2, (226 + 133) x .02 $\cong$ (2 + 5) and (226 + 133 + 184) x .05 $\cong$ (2 + 5 + 20). The actual counts observed in the data are given in Column 6. Thus when (b - d) is over 4.6 logits we expect about two improbable responses and we observe three. When (b - d) is between 3.9 and 4.6 we expect about five improbable responses and again we observe three. Finally when (b - d) is between 2.9 and 3.8 we expect about twenty improbable responses but observe only eight. These data seem to fit the model rather well.

When we scan the 14 most unexpected item and person responses given in Table 5.11.2, we see that they are well dispersed over items and persons. Only Items 3 and 7 and Persons 49M and 95M appear twice and the sexes are equally represented. We must conclude that no clear sign of systematic misfit has been detected in these data.

Nevertheless, in order to use the KCTB example to show the application of item quality control, we will proceed with a further analysis of the six most unexpected responses. These responses of Persons 49M, 68F, 79M, 83F, 93F and 95M to Items 4, 7, 8, 9, 10 and 12 are given in Table 5.11.3. For each of these unexpected incorrect responses, given by able persons on easy items, we have entered the appropriate $(b_v - d_i)$. We have also given for each item its characteristics on the KCT variable, namely its number of taps, reverses and distance and the demographic characteristics of sex, age and grade for each person.

TABLE 5.11.3
THE SIX MOST UNEXPECTED RESPONSES ON KCTB
(101 Persons By 23 Items)

Person Ability	Item Difficulty d_i						Person Characteristics			
b_v	-4.3	-4.1	-2.6	-2.1	-2.1	-0.1	Name	Sex	Age	Grade
0.4	0(4.7)*	1	1	1	1	1	49	M	16+	12+
1.4	1	0(5.5)	1	1	1	1	68	F	16+	12+
1.9	1	1	0(4.5)	1	1	1	79	M	16+	12+
2.4	1	1	1	1	0(4.5)	1	83	F	16+	12+
3.6	1	1	1	0(5.7)	1	1	93	F	16+	12+
4.3	1	1	1	1	1	0(4.4)	95	M	16+	12+

Item Characteristics						
Name	#7	#4	#8	#9	#10	#12
Taps	4	3	4	4	4	4
Reverses	0	0	1	2	2	2
Distance	3	3	5	6	5	7

* $(b - d) = (0.4) - (-4.3) = 4.7$

The difficulty characteristics of the items in reverses and distance show the increase we would expect as the items become more difficult. All six items are on the easy end of the variable. The six persons, on the other hand, are all relatively able adults. This suggests that, if a systematic source of misfit has been detected here, it could only be a slight tendency towards carelessness, or lapses of attention, among some older persons working on items rather too easy for them.

Fit analysis matrices, like Table 5.11.3, which bring together the person and item characteristics of the most unexpected responses, are convenient for supervising the quality of item functioning. These matrices identify and suggest corrections for the systematic sources of item failure shown in the data.

The calculations necessary to evaluate unexpected responses can be accomplished in three ways. The first two are UCON by computer and the hand method explained in Chapter 4. The third way is a crude, but quick, method which often suffices in practical work.

This crude method of fit analysis consists of identifying and calculating only the few largest z^2's observed on an item and then adding to them a 1 for each other person taking that item. This assumes that all of the disturbance observed in that item is due to its outstanding residuals and that the rest of the pattern is more or less as expected.

Table 5.11.4 gives an illustration of this method. There we have taken from Table 5.11.3 just the single largest z_{vi}^2 observed in our KCTB data and added to it a 1 for each other person taking that item, in this case 100. This gives $z_{vi}^2 + 100 = \chi^2$ as the chi-square for that item and $v_i = \chi^2/100$ as the item mean square.

To see whether this crude method can be useful, we will compare it with UCON and the hand method described in Chapter 4, but now applied to these KCTB data. In those procedures we sum all 101 actual z^2's to make our item fit analysis and then divide this sum of squares by its 100 degrees of freedom to get the mean squares shown in Table 5.11.5.

TABLE 5.11.4
CRUDE FIT ANALYSIS FOR SIX KCTB ITEMS

Item Name	Ability minus Difficulty Difference	Single Item z^2	Crude Fit Statistics: Chi-Square $\chi^2 = (z^2 + 100)$	Crude Fit Statistics: Mean Square $v = \chi^2/100$
9	5.7	299	399	4.0
4	5.5	245	345	3.5
7	4.7	110	210	2.1
10	4.5	90	190	1.9
8	4.5	90	190	1.9
12	4.4	81	181	1.8

TABLE 5.11.5
A COMPARISON OF ITEM QUALITY CONTROL METHODS APPLIED TO KCTB

Item Name	UCON Mean Square	Hand Fit Mean Square	Crude Fit Mean Square
9	3.42	3.16	4.0
4	2.64	2.56	3.5
7	1.48	1.40	2.1
10	1.46	1.20	1.9
8	1.43	1.12	1.9
12	1.36	0.98	1.8

The UCON and hand fit methods approximate one another rather closely. Although the crude fit mean squares are somewhat larger in magnitude, their order is identical to the other methods and their values are sufficiently close to get a clear idea concerning the relative fit of these six items. Table 5.11.5 suggests that the crude method can be useful for the quick analysis of item functioning.

5.12 NORM REFERENCING THE KCT VARIABLE

While norms are no more fundamental to the calibration of item banks than are distributions of person heights to the ruling of yardsticks, it is usually useful to know various demographic characteristics of a variable defined by an item bank. Some of these demographic characteristics may even have normative implications under particular circumstances. Because of a shift in emphasis, norming a variable in the Rasch approach takes much less data than norming a test. We need only use enough items to estimate the desired "norming" statistics. Once the variable is normed, then all possible scores from all possible tests drawn from the calibrated bank are automatically norm-referenced through the variable.

Often we are satisfied with a mean and standard deviation for each cell in our normative sampling plan. These two statistics could be estimated from a random sample of 100 or so persons taking a norming test of only two items. Of course, a somewhat longer test of 10 or 15 items will do a better job. Not only will the estimates be better but the extra items will yield standard errors around the norming statistics and thus a test of fit for the plausibility of the data. More than 15 items in a norming test, however, will seldom be necessary. This means that we could norm six different variables simultaneously by allocating 15 items to each of six subtests administered as one 90-item composite test.

We can estimate quick norms from frequency data on bank calibrated items without scoring or measuring the individual persons. This may be useful when trimming sample data is undesirable. If we seek a probability sample from a population, for example, we would rather not distort the sample's status by eliminating some of the persons sampled because they earned zero or perfect scores.

This norming procedure can be accomplished by working directly from the model and the observed number of right answers to each calibrated item.

1. For each sampling cell in the norming study, select from the item bank a suitable set of K calibrated items sufficiently spaced in difficulty d_i to cover the expected ability dispersion of that particular sampling cell. Note that each sampling cell, in principle, has its own individually tailored norming test.
2. Administer this test of K items to a random sample of N persons from the specified cell.
3. Observe the number of persons s_i succeeding on each item.
4. Calculate the natural log odds h_i of these correct answers s_i for each item

$$h_i = \ell n\ [s_i/(N - s_i)] \qquad i = 1,K \qquad \textbf{[5.12.1]}$$

5. Regress these log odds h_i on the associated item difficulties d_i over the K items to obtain the intercept A and slope C of the least squares straight line.
6. Estimate the population mean M and standard deviation SD of that cell's abilities as

$$M = -\ A/C \qquad \textbf{[5.12.2]}$$

$$SD = 1.7\ [(1 - C^2)/C^2]^{1/2} \quad . \qquad \textbf{[5.12.3]}$$

We will apply these procedures to the KCTB sample of 101 persons to see how well they recover the sample mean and standard deviation that we have already estimated from the measurements of each of the 101 persons to be $M' = 0.19$ and $SD' = 2.44$.

We have the item difficulties d_i for Items 3 through 25 and so need only to compute the natural log odds h_i of observed correct answers to each of these items. The values of h_i are given in Column 3 of Table 5.12.1 with the corresponding item difficulties in Column 4.

Regressing these log odds right answers on the item difficulties over the 23 items gives us an intercept of A = 0.07 and a slope of C = -0.56.

Equation 5.12.2 estimates the sample mean as

$$M = -A/C$$

$$= -0.07/-0.56$$

$$= 0.13.$$

TABLE 5.12.1
KCTB LOG ODDS CORRECT ANSWERS AND ITEM DIFFICULTIES

1 Item Name	2 Persons Succeeding s_i	3 Log Odds Correct $h_i=\ell n[s_i/N - s_i)]$	4 Item Difficulty d_i
3	98	3.49	-6.20
4	91	2.21	-4.11
5	82	1.46	-2.58
6	83	1.53	-2.76
7	92	2.32	-4.34
8	82	1.46	-2.58
9	78	1.22	-2.06
10	78	1.22	-2.06
11	68	0.72	-1.03
12	57	0.26	-0.12
13	66	0.63	-0.85
14	62	0.46	-0.52
15	73	0.96	-1.51
16	65	0.59	-0.77
17	30	-0.86	1.93
18	37	-0.55	1.36
19	29	-0.91	2.01
20	20	-1.40	2.88
21	16	-1.67	3.33
22	16	-1.67	3.33
23	8	-2.45	4.52
24	2	-3.90	6.27
25	3	-3.49	5.81
	N = 101	Mean	0.00
		Standard Deviation	3.32

Equation 5.12.3 estimates the sample standard deviation as

$$SD = 1.7[(1 - C^2)/C^2]^{1/2}$$

$$= 1.7\,[(1 - 0.31)/0.31]^{1/2}$$

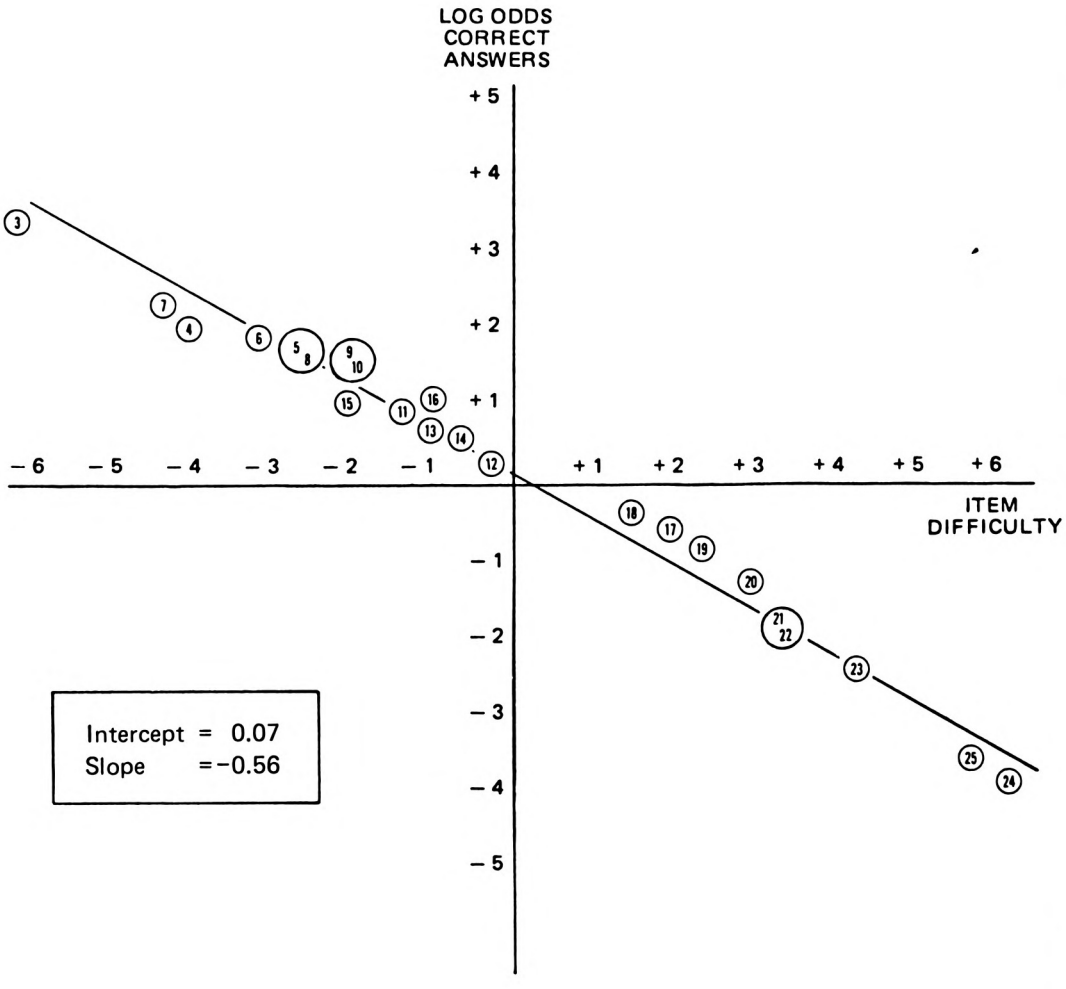
$$= 2.54.$$

These quick norm regression estimates of 0.13 for the mean and 2.54 for the standard deviation compare satisfactorily with the values of 0.19 and 2.44 computed by measuring each of the 101 persons and then calculating their mean and standard deviation in the usual way.

The plot of the log odds correct answers h_i against the item difficulties d_i in Figure 5.12.1 shows how well these norming data fit the straight line expected by the model.

FIGURE 5.12.1

THE QUICK NORMING METHOD OF ESTIMATING SAMPLE MEANS AND STANDARD DEVIATIONS APPLIED TO KCTB



6 DESIGNING TESTS

6.1 INTRODUCTION

In Chapter 5 we have shown how to establish the operational definition of a variable by means of a calibrated bank of items. The next step is to find out how to use these calibrated items to make measures. To do this we must consider two related questions. First, we need to find out how to make the best possible selection of calibrated items from our bank in order to make any particular measurements we have in mind most effective. Second, given such a selection of items and an observed pattern of responses to them, we need to find out how to evaluate the quality of this observation and, if it is valid, how to extract from it the measure we seek, together with its standard error. It is the first question, best test design, which is the major topic of this chapter. Chapter 7 deals with making measures.

6.2 THE MEASUREMENT TARGET

When we plan a measurement, there must be a target person or group of persons about whom we want to know more than we already know. If we care about the quality of our proposed measurements, then we will want to construct our measuring instrument with the specifics of this target in mind. In order to do this systematically we must begin by setting out as clearly as we can what we expect of our target. Where do we suppose it is located on the variable? How uncertain are we of that approximate location? What is the lowest ability we imagine the target could have? What is the highest? How are other possible values distributed in between?

Sometimes we have explicit prior knowledge about our target. We, or others, have measured it before and so we can suggest its probable location and dispersion directly in terms of these prior measures on the variable and their standard errors. Sometimes we can use items calibrated along the variable, some of which we believe are probably just right for the target, some of which are nearly too hard and some of which are nearly too easy. Then we can take from the difficulties of these reference items rough indications of the probable center and boundaries of our target.

One way or another we assemble and clarify our suppositions about our target as well as we can so that we can derive from them the test design which has the best chance of most increasing our knowledge.

Obviously if we know everthing we want to know about our target, then we would not have to measure it in the first place. However, no matter how little we know, we always have some idea of where our target is. Being as clear as possible about that prior knowledge is essential for the design of the best possible test.

Graham A. Douglas collaborated in the preparation of parts of this chapter. See Wright and Douglas, 1975a.

A target specification is a statement about where on the variable we suppose the target to be. We express our best guess by specifying the target's supposed center, its supposed dispersion and perhaps its supposed shape or distribution. If we let

M = our best guess as to target location,
S = our best guess as to target dispersion,
D = our best guess as to target distribution,

then we can describe a target G by the expression G(M,S,D) and we can summarize our prior knowledge, and hence our measurement requirements for any target we wish to measure, by guessing, as well as we can, values for the three target parameters M,S and D.

A picture of a target is given in Figure 6.2.1. Guessing the supposed location M of a target is perfectly straightforward. However, guessing the dispersion S and the distribution D forces us to think through the difference between a target which is a single person and one which is a group. For the single person, S can describe the extent of our uncertainty about where that person is located. The larger our uncertainty, the larger S.

If we can specify boundaries within which we feel fairly sure that the person will be found, we can set S so that M±kS defines these boundaries. Then, even if we have no clear idea at all about the distribution D of our uncertainty between these boundaries, we can nevertheless expect that at least $(1-1/k^2)$ of the possible measures will fall within M±kS.

If we go further and expect that the measures we think possible for the person will pile up near M, then we may even be willing to take a normal distribution as a useful way to describe the shape of our uncertainty. In that case we can expect .95 of the possible measures to fall within M±2S and virtually all of them to fall within M±3S.

We will refer to these two target distributions as the Tchebycheff *interval* and the *normal.* We might consider other target distributions, but these two seem to cover all reasonable target shapes rather well. For example, if we feel unhappy about thinking of our target as approximately normal, then it is unlikely that we will have any definite alternative clearly in mind. Thus, the most likely alternative to a normal target is one of unknown distribution, best captured by a Tchebycheff interval. This realization that all possible target shapes can be satisfactorily represented by just two reasonable alternatives is important because it makes a unique solution to the problem of best test design not only possible but even practical.

If the target is a group rather than an individual, then we may take S and D to be our best guess as to the standard deviation and distribution of that group. If we think the group has a more or less normal distribution, then we will take that as our best guess for D. Otherwise we can always fall back on the Tchebycheff interval.

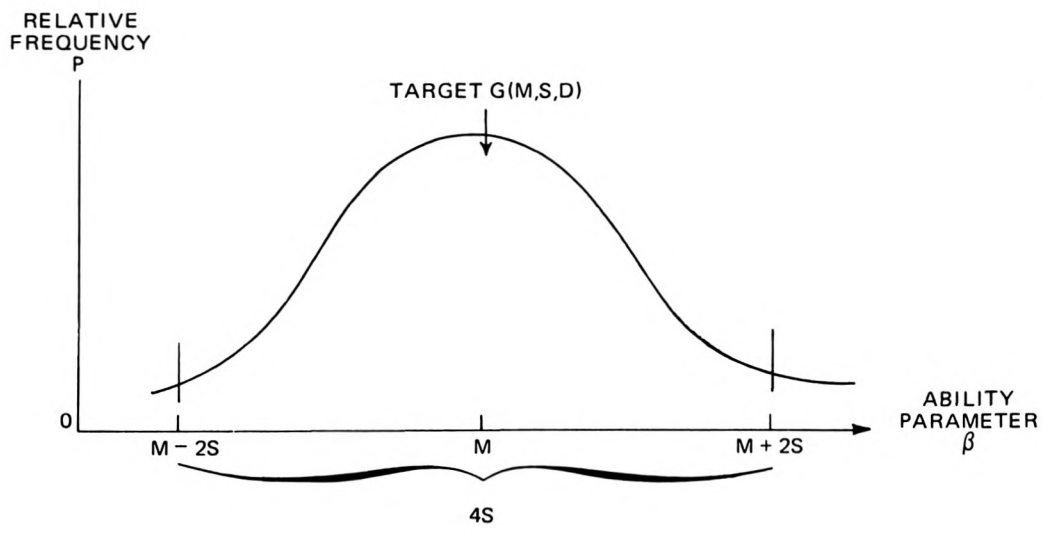
Finally, we must be explicit about how precise we want our measurement to be. After all, this is our motive for measuring. It is just because our present knowledge about our target is too approximate to suit us that we want to know more precisely where our target is and, if it is a group rather than an individual, more precisely about its dispersion. However, whether the target is an individual or a group, our decision about the desired standard error of measurement SEM will be made in terms of individuals, for that, in the end, is what we actually measure.

In the case of a one-person target, we want the SEM to be enough smaller than S to reward our measurement efforts with a useful increase in the precision of our knowledge about where that target person is located. In the case of a group target we want to achieve

an improved estimate not only of M, the center of the group, but also of S, its dispersion. The observable variance of measures over the group estimates not only the underlying variance in ability S^2 but also the measurement error variance SEM^2. Our ability to see the dispersion of our target against the background of measurement error depends on our ability to distinguish between these two components of variance. Since they enter into the observable variance of estimated measures equally, the smaller SEM^2 is with respect to S^2, the more clearly we can identify and estimate S^2, the component due to target dispersion. Thus, for all targets we seek an SEM considerably smaller than S.

FIGURE 6.2.1

THE PICTURE OF A TARGET



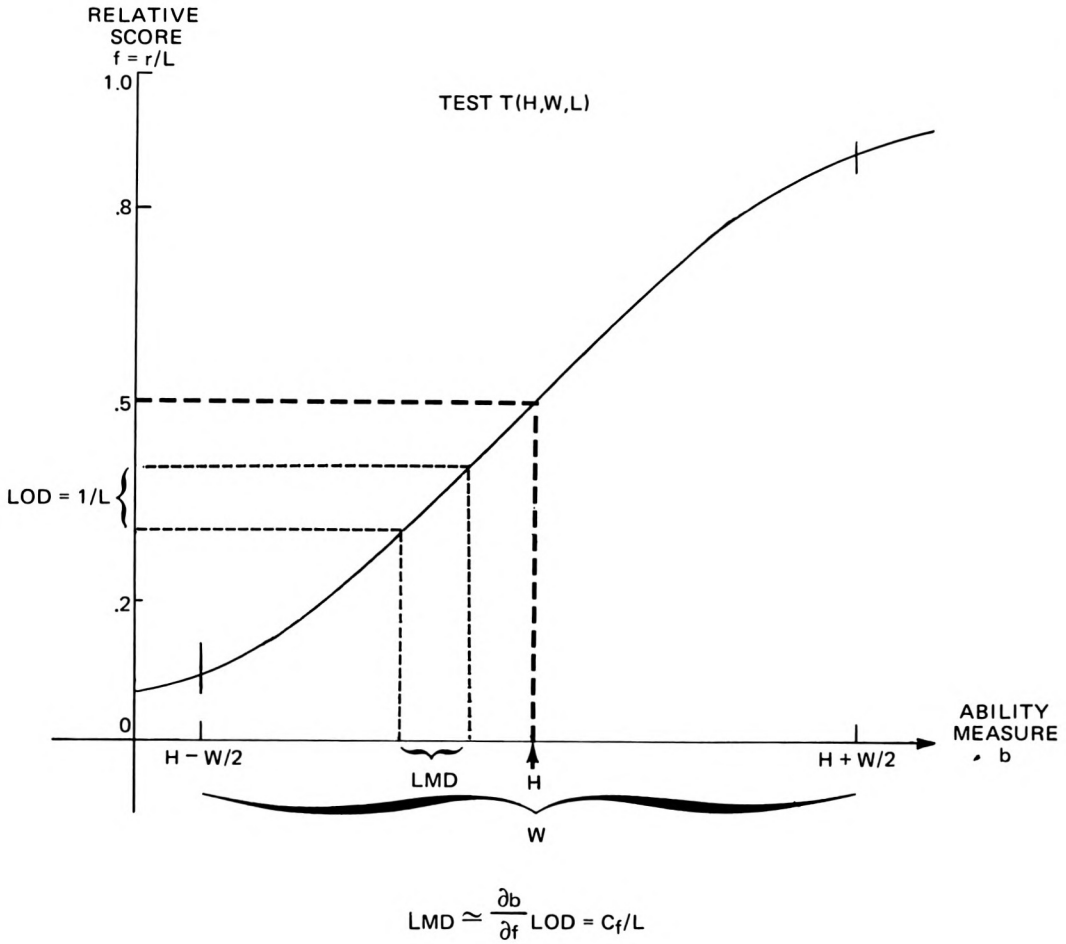
SHAPE D	M ±2S	M ±3S
Interval	.75+	.89+
Normal	.95	.99

6.3 THE MEASURING TEST

A test is a set of suitably calibrated items chosen to go together to form a measuring instrument. The complete specification of a test is the set of all parameters which characterize these items. But when we examine a picture of how a test works to transform observed scores into estimated measures, we see that the operating curve is rather simple and lends itself to specification through just a few test parameters. When the way our items operate fits the Rasch model, then we know that the only item parameters which we need to consider in order to determine the operating characteristics of a test are its item difficulties. When we impose a reasonable fixed distribution on these difficulties, then no matter how many items we use, we can reduce the number of test parameters to only three.

FIGURE 6.3.1

THE OPERATION OF A TEST



In Figure 6.3.1 we can see from the shape of the test operating curve that its two outstanding features are its position along the variable, which we will call test *height*, and the range of abilities over which the test can measure more or less accurately, a characteristic caused primarily by the dispersion of the item difficulties, which we will call test *width*.

But height and width do not complete the characterization of a test. When we look more closely at the way the test curve transforms observed scores into inferred measures we see that there is a discontinuity in observable scores which is going to determine the smallest increment in ability we can measure with any particular test. This least measurable difference LMD depends on the test's least observable difference LOD. Since the least change possible in a test score is one, the LOD in relative score f = r/L, must be 1/L. In Section 6.5 we will find that the standard error of measurement, or least believable difference, SEM also depends on the number of items in the test. Indeed SEM = LMD^{½}. So in order to finish characterizing a test we must also specify its *length*.

From this we see that any test design can be defined more or less completely just by specifying the three test characteristics; height, width and length. If we let

H = the height of the test on the variable, that is, the average difficulty of its selected items,
W = the width of the test in item difficulties, that is, the range of its item difficulties and
L = the length of the test in number of items,

then we can describe a test design T by the convenient expression T(H,W,L).

In the practical application of best test design, however, we will have to approximate our best design T for a target G from a finite pool of existing items. In order to discriminate in our thinking between the best test design T(H,W,L) and its approximate realization in practice, we will describe an actual test as t(h,w,L) where

h = the average difficulty of its actual items, and
w = an estimate of their actual difficulty range.

6.4 THE SHAPE OF A BEST TEST

A best test is one which measures best in the region within which measurements are expected to occur.* Measuring best means measuring most precisely. A best test design T(H,W,L) is one with the smallest error of measurement SEM over the target G(M,S,D) for given length L (or, what is equivalent, with the smallest L for a given value of SEM). "Over the target" implies the minimization of a distribution of possible SEMs. Thus, a position with respect to the most likely target distribution must be taken before the minimization of SEM can proceed.

We bring the profusion of possible target shapes under control by focusing on the two extremes—interval and normal. How shall minimization be specified in each case? For a normal target it seems reasonable to maximize average precision, that is, to minimize average SEM, over the whole target.

To decide what to do for an interval target, we need to know how the SEM varies over possible test scores. When we derive an exact form for the precision of measurement, we find that for ordinary tests with less than three logits between adjacent items, precision is a maximum for measurements made at the center of the test and decreases as test and target are increasingly off-center with respect to one another. For tests centered on their targets this means that maximizing precision at the boundaries of an interval target is a good way to maximize precision over the target interval. So for interval targets we will maximize precision at the target boundaries.

When we derive the SEM^2 from our response model we will discover that it is the reciprocal of the information about ability supplied by each item response averaged over the test. Since the most informative items are those nearest the ability being measured

*Attempts to meet this requirement have been made by Birnbaum (1968, pp. 465-471). Our ideas are consistent with his efforts, but we have taken them to their logical and practical conclusion.

and the least informative are those farthest away, the precision over the target will depend not only on the distribution of the target but also on the shape of the test. Thus, the question of what is a best test also depends on our taking a position with respect to the best distribution of test item difficulties.

What are the reasonable possibilities? If we want to measure a normal target, then a test made up of normally distributed item difficulties ought to produce the best maximization of precision over the target. This is the conclusion implied in Birnbaum's analysis of information maximization (Birnbaum, 1968, p. 467).

However, normal tests are clumsy to compose. Normal order statistics can be used to define a set of item difficulties, but this is tedious. More problematic is the odd conception of measuring implied by an instrument composed of normally distributed measuring elements. A normal test would be like a yardstick with rulings bunched in the middle and spread at the ends. Measuring with such an irregularly ruled yardstick would be awkward. In the long run, even for normal targets, our interest becomes spread out evenly over all the abilities which might be measured by a test. Equally spaced items are the test shape which serves that interest best. That is the way we construct yardsticks. The test design corresponding to an evenly ruled yardstick is the uniform test in which items are evenly spaced from easiest to hardest (Birnbaum, 1968, p. 466).

Two target distributions, normal and interval, and two test shapes, normal and uniform, produce four possible combinations of target and test. Wright and Douglas (1975a) investigated all four combinations rather extensively and found the normal test to work best on the normal target and the uniform test to work best on the interval target. When they compared the normal and uniform tests on normal targets, however, these two test shapes differed so little in their measuring precision as to appear equivalent for all practical purposes. Thus the best all purpose test shape is the uniform test.

6.5 THE PRECISION OF A BEST TEST

Now we turn to the response model formulation of the standard error of measurement SEM so that we can become explicit about which test designs maximize precision by minimizing SEM. We must find out how the test design T(H,W,L) influences SEM and how we can vary the test characteristics of H, W and L in response to a target specification G(M,S,D) in order to minimize SEM over that target.

The response model specifies

$$p_{fi} = \exp(b_f - d_i)/[1 + \exp(b_f - d_i)] \qquad [6.5.1]$$

where

p_{fi} = the probability of a correct response at f and i,
b_f = the ability estimate at relative score f = r/L,
d_i = the calibrated difficulty of item i.

The measure b_f is estimated from a test of length L with items $\{d_i\}$ for i = 1, L through the equation (for details see Sections 1.5 and 3.7)

$$f = \sum_i^L p_{fi} / L, \qquad \text{for } f = 1/L,(L-1)/L \qquad [6.5.2]$$

with asymptotic variance

$$1 / \sum_{i}^{L} p_{fi} (1 - p_{fi}) = SEM_f^2 \qquad [6.5.3]$$

This is the square of the standard error of measurement at relative score f.

We see that SEM_f depends on the sum of $p_{fi} (1 - p_{fi})$ over i. Thus it is a function of b_f and all the d_i. However, fluctuations in p (1 - p) are rather mild for p between 0.2 and 0.8. To expedite insight into the make-up of SEM_f we can reformulate it so that the average value of $p_{fi} (1 - p_{fi})$ over i is one component and test length L is the other.

$$SEM_f = \left\{ L / [\sum_{i}^{L} p_{fi} (1 - p_{fi})] \right\}^{1/2} (1/L)^{1/2} = (C_f/L)^{1/2} \qquad [6.5.4]$$

in which

$$C_f = [\sum_{i}^{L} p_{fi} (1 - p_{fi})/L]^{-1}$$

In this expression we factor test length L out of SEM in order to find a length-free error coefficient C_f.

Resuming our study of the operating curve of a test given in Figure 6.3.1 we see that the least measurable difference in ability LMD is $(\frac{\partial b}{\partial f})$ LOD. Since the least observable increment in relative score is 1/L, all we need to complete the formulation of the LMD is the derivative of b with respect to f which from Equations 6.5.1 and 6.5.2 is

$$\frac{\partial b}{\partial f} = [\sum_{i}^{L} p_{fi} (1 - p_{fi})/L]^{-1} \qquad [6.5.5.]$$

But this is our error coefficient C_f, thus the least measurable difference at relative score f is

$$LMD_f \simeq C_f/L \qquad [6.5.6]$$

and

$$SEM_f = LMD_f^{1/2} \simeq (C_f/L)^{1/2}$$

With SEM_f in this form we note that, as far as test shape is concerned, it is C_f which requires minimization. This will be true whether we use C_{min} to minimize SEM_f given L or to minimize L given SEM_f.

6.6 THE ERROR COEFFICIENT

Now we need to know more about this error coefficient C_f. The essential ingredient of C_f is the expression $p_{fi} (1 - p_{fi})$. This is the information I_{fi} on b_f contained in a response to item i with difficulty d_i (Birnbaum, 1968, p. 460-68). Its average value

$$I_{f.} = \sum_{i}^{L} I_{fi}/L = C_f^{-1}$$

over the items on a test is the average information about b_f per item provided by that test. Thus C_f is the reciprocal of average test information. The greater the information obtained by a test the smaller C_f and hence the smaller SEM_f and so the greater the precision.

What values can we expect C_f to take? We can approach this question in two ways: in terms of the influence of reasonable values of $(b_f - d_i)$ on p_{fi} and, for uniform tests, in terms of test width W and the boundary probabilities p_{f1} for i = 1, the easiest item, and p_{fL} for i = L, the hardest item. The probability p_{fi} is defined in Equation 6.5.1.

Beginning with reasonable values of $(b_f - d_i)$, we see that when $b_f = d_i$ and their difference is zero, then $p_{fi} = 1/2$, $p_{fi}(1 - p_{fi}) = 1/4$ and $C_f = 4$, but when $(b_f - d_i) = -2$ then $p_{fi} = 1/8$, $p_{fi}(1 - p_{fi}) = 1/9$ and $C_f = 9$. (Notice that $C_f = 9$ when $(b_f - d_i) = +2$ and $p_{fi} = 7/8$ also). Since an average can never be greater than its maximum element nor less than its minimum, we can use these figures as bounds for C_f.

When $-2 < (b_f - d_i) < +2$

then $1/8 < p_{fi} < 7/8$

and $4 < C_f < 9.$ **[6.6.1]**

Turning to the bounds we can derive for C_f from the test width W and the boundary probabilities p_{f1} and p_{fL} of a uniform test, we can use an expression for C_f given W derived in Wright and Douglas, 1975a (also Birnbaum, 1968, p. 466).

$$C_{fw} = W/(p_{f1} - p_{fL})$$

where

W = the item difficulty width of a uniform test,

p_{f1} = the probability of a correct response by b_f to the easiest item on the test, and

p_{fL} = the probability of a correct response by b_f to the hardest item on the test.

When b_f is contained within the difficulty boundaries of the test, and W is greater than 4 then $1/2 < (p_{f1} - p_{fL}) < 1$ and C_{fw} must fall between W and 2W, that is

When $d_1 < b_f < d_L$

and $W > 4$

then $W < C_{fw} < 2W.$ **[6.6.2]**

It follows from these considerations that SEM = $(C/L)^{1/2}$ is bounded by

$$2/L^{1/2} < SEM < 3/L^{1/2}$$

for any test on which

$$-2 < (b_f - d_i) < +2,$$

and by

$$(W/L)^{1/2} < SEM < (2W/L)^{1/2}$$

for uniform tests when

$$W > 4 \text{ and } d_1 < b_f < d_L.$$

6.7 THE DESIGN OF A BEST TEST

Best test design depends on relating the characteristics of test design T(H,W,L) to the characteristics of target G(M,S,D) so that the SEM is minimized in the region of the variable where the measurements are expected to take place. The relationship between test and target visible in Figure 6.7.1 makes the general principles of best test design obvious. To match test to target we aim the height of the test at the center of the target, widen the test sufficiently to cover target dispersion and lengthen the test until it provides the precision we require.

For best test design on either interval or normal targets we select a set of equivalent items (where W = 0) or a set of uniform items with the W indicated in Table 6.7.1. Table 6.7.1 gives optimal uniform test widths for normal and interval targets. For example, if the target is thought to be approximately normal with presumed standard deviation S = 1.5, the optimum test width W is 4. If, however, the target is more uniform in shape then the optimum width could be as large as 8. Note that for any value of S a smaller W is always indicated when a normal "bunched up" target shape is expected.

Table 6.7.1 also shows the efficiency of a simple rule for relating test width W to target dispersion S. The rule W = 4S comes close to the optimum W for narrow interval targets and for wide normal targets. When we are vague about where our target is we are also vague about its boundaries. That is just the situation where we would be willing to use a normal distribution as the shape of our target uncertainty. When our target is narrow however, that is the time when we are rather sure of our target boundaries but, perhaps, not so willing to specify our expectations as to its precise distribution within these narrow boundaries. To the extent that interval shapes are natural for narrow targets while normal shapes are inevitable for wide targets, W = 4S is a useful simple rule.

The efficiency of this simple rule for normal and interval targets is given in the final columns of Table 6.7.1. There we see that its efficiency is hardly ever less than 90 per cent. If we cross over from an interval target to a normal target as our expected target dispersion exceeds 1.4, then the efficiency is never less than 95 per cent. This means, for example, that a simple rule test of 20 items is never less precise than an optimum test of 19 items.

Our investigations have shown that given a target M, S and D there exists an optimum test design H and W from which we may generate a unique set of L uniformly distributed item parameters $\{\delta_i\}$. However, this design is an idealization and cannot be perfected in practice. Real item banks are finite and each item difficulty is only an estimate of its corresponding parameter and hence inevitably subject to calibration error. We will never be able to select the exact items stipulated by the best test design $\{\delta_i\}$. Instead we must attempt to select among the items available, a real set of $\{d_i\}$ which comes as close as possible to our ideal design $\{\delta_i\}$.

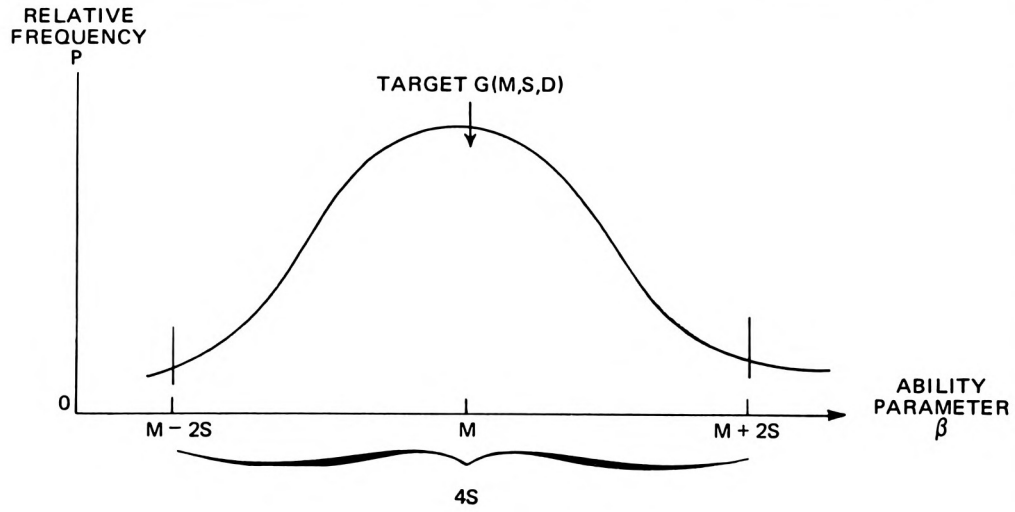
Thus parallel to the design specification T(H,W,L) we must write the test description t(h,w,L) characterizing the actual test $\{d_i\}$ which we can construct in practice. This raises the problem of estimating h and w.

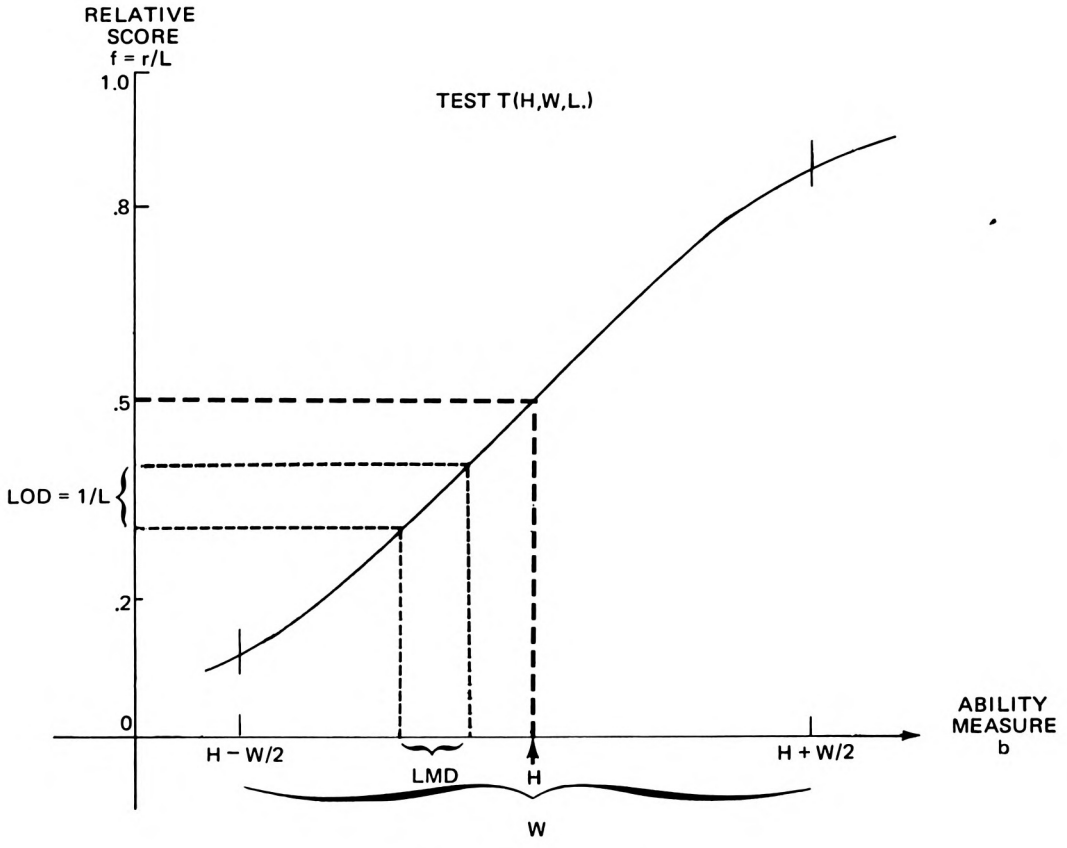
The estimated test height h can be determined by the average estimated difficulties of the test items

$$h = \sum_{i}^{L} d_i/L = d. \qquad [6.7.1]$$

FIGURE 6.7.1

DISTRIBUTION OF A TARGET AND OPERATION OF A TEST





BEST TEST DESIGN
H = M
W = 4S
$L = C_f/SEM^2$
$LMD = C_f/L$
$4 < C_f < 9$

The estimated test width w can be determined from the range of these estimated difficulties, or perhaps a bit more precisely from an estimate of this range based on the two easiest items d_1 and d_2, and the two hardest, d_{L-1} and d_L.

$$w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)] \qquad [6.7.2]$$

TABLE 6.7.1

OPTIMUM VALUES OF W FOR BEST UNIFORM TESTS ON NORMAL AND INTERVAL TARGETS

TARGET STD. DEV. S	NORMAL TARGET error minimized over $N(M,S^2)$	INTERVAL TARGET error minimized at $(M \pm 2S)$	SIMPLE RULE* W=4S	EFFICIENCY** Normal	EFFICIENCY** Interval
.5	0	0	2	94	97
.6	0	0	2		
.7	0	2	3	90	100
.8	0	3	3		
.9	0	4	4		
1.0	0	5	4	89	98
1.1	0	6	4		
1.2	1	6	5	92	96
1.3	2	7	5		
1.4	3	7	6		
1.5	4	8	6	96	91
1.6	5	9	6		
1.8	6	10	7	98	87
2.0	8	11	8	99	84

*This Simple Rule is conservative for narrow targets and more practical since available items are bound to spread some. It is also close to the normal target optimum for wide targets, which is reasonable in the face of substantial target uncertainty.

**Efficiency = $C_W / C_{4S} = L_W / L_{4S}$

where C_W = minimum error coefficient for optimum W.
C_{4S} = error coefficient for W = 4S.
L_W = length of optimum test of width W.
L_{4S} = length of equally precise test of width 4S.

6.8 THE COMPLETE RULES FOR BEST TEST DESIGN

We are now in a position to give explicit, objective and systematic rules for the design and use of a best possible test. To design test T(H,W,L) for target G(M,S,D):

1. From our hypothesis about M we derive H = M.
2. From our hypothesis about S we derive an optimum W either by consulting Table 6.7.1 or by using the simple rule W = 4S.

3. From our requirements for the measurement precision SEM we seek, we derive $L = C/SEM^2$. A fairly accurate value for C can be found in Table 6.8.1 which gives, at various expected relative scores, f the value of C minimized by the W chosen in Step 2. Alternatively C can be approximated by one of the simple rules C = 6S or C = 6 from Equations 6.6.1 or 6.6.2.
4. From these H, W and L we generate the design set of items $\{\delta_i\}$ according to the formula

$$\delta_i = H - (W/2)[(L - 2i + 1)/L] \qquad \text{for } i = 1,L$$

Then for test t(h,w,L) from design T(H,W,L)

5. We select items d_i from our item bank such that they best approximate the set $\{\delta_i\}$ by minimizing the discrepancy $(d_i - \delta_i)$.
6. We calculate

$$h = \sum_i^L d_i/L = d.$$

and

$$w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$$

7. We administer the set of $\{d_i\}$ as the test t(h,w,L).

TABLE 6.8.1

ERROR COEFFICIENT C_{fW} FOR SELECTED TEST WIDTH W AND EXPECTED RELATIVE SCORE f FOR UNIFORM TESTS

Expected Relative Score f	Test Width W					
	0	2	4	6	8	10
.10	10.9	11.6	13.0	13.7	15.2	16.0
.20	6.3	6.8	7.3	8.4	10.2	11.6
.30	4.8	5.3	5.8	7.3	9.0	10.2
.40	4.0	4.4	5.3	6.8	8.4	10.2
.50	4.0	4.4	5.3	6.8	8.4	10.2
.60	4.0	4.4	5.3	6.8	8.4	10.2
.70	4.8	5.3	5.8	7.3	9.0	10.2
.80	6.3	6.8	7.3	8.4	10.2	11.6
.90	10.9	11.6	13.0	13.7	15.2	16.0

$$C_{fW} = W\,[1 - \exp(-W)] \,/\, \{[1 - \exp(-fW)]\,[1 - \exp(-(1-f)\,W)]\}$$

7 MAKING MEASURES

7.1 USING A VARIABLE TO MAKE MEASURES

This chapter is about turning test scores into measures. But before we show how to do this in Sections 7.2 and 7.3, we will review how the test items defining a variable can be used to make measures.

To make a measure we collect and combine a series of observed responses in such a way that they support an inference as to the position of the person on a variable. We summarize these observations into a score based on them and this score is used to imply the measure of the person on the variable. The variable itself, however, is an idea and not a direct experience. Its nature can only be inferred from relevant samples of carefully selected observations.

The purpose of a variable is to provide a basis for comparing persons and generalizing about their relative status. This purpose requires the achievement of objectivity in the variable's definition and in the way measures on it are made. The idea of the variable transcends any particular set of observations and the measure on the variable must transcend the observed responses on which it is based. Making measures with tests requires objectively calibrated test items which provoke the observed item responses and then through their calibrations carry these responses onto the scale of the variable. It is these items that operationally define the variable and bring meaning to the measurement of the person.

Different ways of getting a particular score on a test do not generally arouse different opinions of the abilities of persons taking the test. When two persons earn the same score, we seldom put one person ahead of the other because they answered particular items successfully. This is because we think of each score as resulting from the same exposure to the same items giving each person's ability the same opportunity to express itself. But whenever we are willing to take identical scores to have equivalent meaning and do not care which items are actually answered correctly we are practicing "item-free" measurement. This widespread practice of item-free measurement within a test implies, without further assumption, test-free measurement within a bank of calibrated items.

A calibrated item bank provides a resource from which subsets of items can be selected to form specifically designed tests with optimal characteristics. Scores on these tests, although stemming from different combinations of "correct" responses to different selections of items, can nevertheless be converted through the bank calibrations into comparable measures. Procedures for obtaining comparable measures for individualized tests are given in Sections 7.4 to 7.7.

To validate these measures, however, we must assess the extent to which the persons in question have taken the items in the way we intended them to be taken. The item calibrations in the bank come from occasions on which many persons were found to respond to these items in a particular consistent way. This is the context in which the item calibrations gained their meaning. The meaning these calibrations now convey depends on how the new persons being measured are found to respond to the items. The validity of

their measures depends on the presence of acceptable relations between what we actually observe and what we expect to observe according to our measurement model and our item calibrations. Thus before we can accept any measure as valid, we must examine the plausibility of the pattern of responses on which that measure is based. The procedure for accomplishing the analysis of person fit necessary to establish measure validity is given in Sections 7.8 and 7.9. In these sections we show how to detect person misfit and what various kinds of misfit look like.

Whenever misfit is identified, the next step is to deal with the measurement quality control problem this misfit causes. If we can identify the circumstances leading to the misfit, we may be able to extract from the flawed response record a measure which the observed pattern of responses can sustain. We show how to do this in Section 7.10.

7.2 CONVERTING SCORES TO MEASURES BY UCON, PROX AND UFORM

When a person takes a test, the resulting observation of the person is their test score. To see how to get from this test score r to the estimated measure b which it implies we refer to the measurement model,

$$P\{x_i = 1\} = \pi_i = \exp(\beta - \delta_i)/[1 + \exp(\beta - \delta_i)] \qquad [7.2.1]$$

which specifies how item calibration δ_i and person measure β are implied by the person's observed response x_i. The model implies that for each response of a person to an item we "expect" an intermediate "probable" value which is neither $x_i = 1$ for a correct response nor $x_i = 0$ for an incorrect response, but somewhere in between them. This "expected" value is the probability π_i given in Equation 7.2.1 that $x_i = 1$, and it works just like our expectation that fair coins fall half the time heads. Since we "expect" a value on each coin toss which is half the time heads and half the time tails, even though what happens can only be one or the other, our expected value for a particular toss is neither 0 nor 1, but half way between at $\pi = ½$.

Thus the expected value of response x_i is

$$E\{x_i\} = \pi_i$$

the model probability of a correct answer to item i.

Since the test score $r = \sum_i^L x_i$ is the sum of the item responses, the expected value of r is the sum of their expectations,

$$E\{r\} = E\left\{\sum_i^L x_i\right\} = \sum_i^L E\{x_i\} = \sum_i^L \pi_i \quad .$$

If we now substitute in π_i the measure b_r to be estimated for β on the basis of score r and the estimated calibrations $\{d_i\}$ for $\{\delta_i\}$, we have an estimation equation which relates r and b_r as follows

$$r = \sum_i^L \exp(b_r - d_i)/[1 + \exp(b_r - d_i)] \quad . \qquad [7.2.2]$$

From this equation, a person's score r and the calibrations $\{d_i\}$ of the items taken, we can determine the measure b_r which they imply.

One way to solve Equation 7.2.2 is to use the UCON procedure described in Chapter 3. The UCON estimated measure is obtained by performing j = 1, m iterations of

$$b_r^{j+1} = b_r^j + (r - \sum_i^L p_{ri}^j)/[\sum_i^L p_{ri}^j(1 - p_{ri}^j)] \qquad [7.2.3]$$

in which

$$p_{ri}^{j} = \exp(b_r^{j} - d_i)/[1 + \exp(b_r^{j} - d_i)] \quad . \qquad [7.2.4]$$

and the first value of b_r^{j} is

$$b_r^{o} = \ell n\,[\,r/(L - r)] \quad .$$

This UCON procedure requires 3 or 4 iterations and a convergence criterion for successive values of b_r^{j} such as

$$|\,b_r^{j+1} - b_r^{j}\,| < .01 \text{ logits} \quad .$$

When the convergence criterion is reached, then the estimated measure is the last value of b_r, namely

$$b_r = b_r^{j+1} \qquad [7.2.5]$$

with standard error

$$s_r = [\sum_{i}^{L} p_{ri}^{j}\,(1 - p_{ri}^{j})]^{-1/2} \qquad [7.2.6]$$

The UCON procedure responds in detail to the distribution of item difficulties $\{d_i\}$ and so estimates a measure b_r which is completely freed of whatever distribution of item difficulties characterizes the test. When the items happen to be such that their d_i's approximate a normal distribution $d_i \sim N(H, \sigma_d^2)$, however, then the PROX procedure described in Chapter 2 is an excellent approximation to the UCON procedure.

The PROX estimated measure b_r can be found without iteration as

$$b_r = h + [1 + (s_d^2/2.89)]^{1/2}\,\ell n\,[\,r/(L - r)] \qquad [7.2.7]$$

in which

$$h = \sum_{i}^{L} d_i/L = d.$$

estimates test height H and

$$s_d^2 = (\sum_{i}^{L} d_i^2 - Ld.^2)/(L - 1)$$

estimates the variance of test item difficulty σ_d^2. The standard error for this b_r is

$$s_r = (1 + s_d^2/2.89)^{1/2}\,[L/r\,(L - r)]^{1/2} \quad . \qquad [7.2.8]$$

Since it is often the case that the d_i's of a sample of new items approximate a normal distribution and since normal samples of persons are typical, PROX is often useful for calibrating new items. In making measures, however, we can take advantage of already calibrated items and spread them uniformly $d_i \sim U(H, W)$ over the range of ability to be measured. Such a uniform test can be described completely by its height H, width W, and length L. Its measures can be calculated efficiently by the UFORM procedure described in Section 7.3.

The UFORM estimated measure b_f is

$$b_f = h + w\,(f - 0.5) + \ell n\,(A/B) \qquad [7.2.9]$$

where $A = 1 - \exp(-wf)$

$B = 1 - \exp[-w(1-f)]$

and $h = \sum_{i}^{L} d_i/L = d.$

estimates test height H,

$$w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$$

estimates test width W, and

$$f = r/L$$

is the relative score on the L item test (Wright and Douglas, 1975a, 21-23).

The standard error for this b_f is

$$s_f = [(w/L)(C/AB)]^{1/2} \qquad [7.2.10]$$

where $A = 1 - \exp(-wf)$

$B = 1 - \exp[-w(1-f)]$

$C = 1 - \exp(-w)$.

To illustrate the use of these procedures we have chosen nine persons from our KCTB sample of 101. Three of these persons are at the preschool level, three are at the primary level and three are adults.

In Columns 2 through 4 of Table 7.2.1 we give the sex, age and grade of these nine persons. Column 5 contains their KCTB scores. Their corresponding UCON abilities are given in Column 6.

TABLE 7.2.1

NINE PERSONS SELECTED FROM KCTB SAMPLE

	1	2	3	4	5	6
Ability Group	Person Name	Sex	Age in Years	School Grade	KCTB Score	UCON Ability
Preschool	3M	M	3	Preschool	1	-5.8
	6F	F	5	Preschool	3	-3.9
	12M	M	4	Preschool	5	-2.8
Primary	29M	M	6	1	10	-0.9
	35F	F	9	4	11	-0.5
	69M	M	8	4	15	1.4
Adult	88M	M	17+	12+	18	3.0
	98F	F	16	11	20	4.3
	101F	F	17+	12+	21	5.2

7.3 MEASURES FROM BEST TESTS BY UFORM

Prior to item calibration our only knowledge of item difficulties comes from our general concept of the variable which the items are supposed to define. We do not know the actual distribution of these items along their variable. Once we have calibrated items, however, as with KCTB, then we have a detailed picture of where these items are located. As a result we can use specially selected subsets of these calibrated items to expedite measurement.

These specially designed or "tailored" tests will vary in length, in difficulty level and in range of ability covered depending on the measurement target. Estimating measures from such subsets of items can be done efficiently because we can construct the distribution of item difficulties to suit our purpose. In particular, if we want to optimize the efficiency of our designed tests, we will construct them so that the items are uniformly spaced in difficulty over their measurement target. This makes the estimation of measures from scores on these tests entirely manageable by the simple UFORM procedure.

All that is needed to apply UFORM are estimates of the height H, width W and length L of a test. Then a single conversion table arranged by relative score and test width provides all the person measures ever needed. A second table, similarly arranged, gives the coefficients necessary to form the standard errors of these measures. Table 7.3.1 is an abbreviated table of these relative measures and Table 7.3.2 is an abbreviated table of their error coefficients. More complete tables for relative measures and their error coefficients are given in Appendix Tables A and B.

To use Tables 7.3.1 and 7.3.2 (or Tables A and B) we need the approximate width w of the test and the person's relative score $f = r/L$. Together they determine the person's relative ability x_{fw} and its corresponding error coefficient C_{fw}. When we combine this information with test height h and test length L, we get the measure $b_{fw} = h + x_{fw}$ and its standard error $s_{fw} = C_{fw}^{1/2}/L^{1/2}$.

In order to use Tables 7.3.1 and 7.3.2 for a particular test, we need estimates of that test's basic characteristics H, W and L. Test length L is self-evident. Test height H is estimated from the average difficulty level of the test's items, namely $h = \sum_i^L d_i/L = d..$ The estimation of test width W, however, can be problematic when an irregular distribution of item difficulties at the extremes of the test cannot be avoided.

Test width can be estimated in various ways. For example

$$w_1 = (d_L - d_1)[L/(L-1)]$$

$$w_2 = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L-2)]$$

$$w_3 = [(d_L + d_{L-1} + d_{L-2} - d_3 - d_2 - d_1)/3][L/(L-3)]$$

or

$$w_s = 3.5s_d \quad \text{where} \quad s_d^2 = (\sum_i^L d_i^2 - Ld.^2)/(L-1) \quad .$$

The method we have found best in practice is w_2, the one based on the average difference between the two easiest and the two hardest items. This procedure for estimating test width is illustrated in Table 7.3.3 where we calculate w for five forms of the KCTB.

TABLE 7.3.1

RELATIVE MEASURE x_{fw} FOR UNIFORM TESTS

Relative Score f	Test Width w 2	4	6	8	10
.1	- 2.3	- 2.7	- 3.2	- 3.8	- 4.5
.2	- 1.5	- 1.8	- 2.2	- 2.6	- 3.1
.3	- 0.9	- 1.1	- 1.4	- 1.7	- 2.1
.4	- 0.4	- 0.5	- 0.7	- 0.8	- 1.0
.5	0.0	0.0	0.0	0.0	0.0
.6	0.4	0.5	0.7	0.8	1.0
.7	0.9	1.1	1.4	1.7	2.1
.8	1.5	1.8	2.2	2.6	3.1
.9	2.3	2.7	3.2	3.8	4.5

For more detail see Appendix Table A

Test Length: L

Relative Score: $f = r/L$

Test Height: $h = \sum_{i}^{L} d_i/L$

Test Width: $w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$

Measure: $b_f = h + x_{fw}$

TABLE 7.3.2

ERROR COEFFICIENT $C^{1/2}{}_{fw}$ FOR UNIFORM TESTS

Relative Score f	Test Width w 2	4	6	8	10
.1	3.4	3.5	3.7	3.8	4.0
.2	2.6	2.7	2.9	3.2	3.4
.3	2.3	2.4	2.7	3.0	3.2
.4	2.1	2.3	2.6	2.9	3.2
.5	2.1	2.3	2.6	2.9	3.2
.6	2.1	2.3	2.6	2.9	3.2
.7	2.3	2.4	2.7	3.0	3.2
.8	2.6	2.7	2.9	3.2	3.4
.9	3.4	3.5	3.7	3.8	4.0

For more detail see Appendix Table B.

Standard Error: $s_{fw} = C^{1/2}{}_{fw}/L^{1/2}$

TABLE 7.3.3

ESTIMATING TEST WIDTH w FOR FIVE KCT FORMS

Test Form	Test Length	Two Easiest Items		Two Hardest Items		Test Width	
						Calculated w′	Used w
KCTB	23	−6.2*	−4.3	5.8	6.3	12.4*	
	22	−4.3	−4.1	5.8	6.3	11.3	11
Preschool	8	−6.2	−4.3	−2.1	−2.1	4.2	4
Primary	15	−2.7	−2.6	2.0	2.9	5.9	6
Adult	15	−1.5	−1.0	5.8	6.3	8.4	8
Pilot	7	−6.2	−4.1	4.5	6.3	14.8	15

$w' = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$

w = w′ rounded to nearest integer.

* Item 3 at −6.2 is 2 logits below the more or less uniform stream of 22 items from Item 7 at −4.3 through Item 24 at 6.3. UFORM is more accurate with this kind of extreme non-uniformity, when test width is calculated without the very irregular extreme item.

The first row of Table 7.3.3 concerns the 23 items in the KCTB "item bank." From these 23 items we have composed three narrow-range test forms focused on three ability levels: a Preschool Form of 8 items, a Primary Form of 15 items and an Adult Form of 15 items, and also one wide-range Pilot Form of 7 items. The calibrations for the two hardest and two easiest items for each of these test forms are given in Table 7.3.3. With these calibrations we can estimate the various test widths using the w_2 method to calculate w′ as

$$w' = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)] \qquad [7.3.1]$$

and rounding the w′ computed to the nearest integer for the value of w used in tables like 7.3.1 and 7.3.2.

Now we apply the UCON, PROX and UFORM measuring procedures to our sample of nine persons and compare the results. Table 7.3.4 gives the UCON measures and errors for the KCTB scores from 1 to 22. Table 7.3.5 gives the nine persons' KCTB scores and the corresponding ability measures and errors for each of these scores by UCON, PROX and UFORM.

Person 29M, for example, earned a KCTB score of 10 correct out of 23 items attempted. His UCON ability and error, looked up in Table 7.3.4, are b = −0.9 and s = 0.6

His PROX ability and error calculated from h = 0, X = 2.2 and L = 23 are

$$b = h + X \ln [r/(L - r)]$$

$$= 0 + 2.2 \ln [10/13]$$

$$= -0.6$$

and

$$s = X[L/r(L - r)]^{1/2}$$

$$= 2.2[23/10\ (13)]^{1/2}$$

$$= 0.9$$

The value of the expansion factor X comes from the variance of item difficulty $s_d^2 = 11.0$ as

$$X = (1 + s_d^2/2.89)^{1/2}$$

$$= (1 + 11/2.89)^{1/2}$$

$$= 2.2 \quad .$$

TABLE 7.3.4
UCON ABILITIES AND ERRORS FOR THE 23 KCTB ITEMS

Score r	Ability b	Error s
1	-5.8	1.2
2	-4.6	1.0
3	-3.9	0.8
4	-3.3	0.7
5	-2.8	0.7
6	-2.4	0.6
7	-2.0	0.6
8	-1.6	0.6
9	-1.3	0.6
10	-0.9	0.6
11	-0.5	0.6
12	-0.1	0.7
13	0.3	0.7
14	0.8	0.7
15	1.4	0.7
16	1.9	0.8
17	2.4	0.8
18	3.0	0.8
19	3.6	0.8
20	4.3	0.9
21	5.2	1.0
22	6.3	1.2

TABLE 7.3.5

A COMPARISON OF KCTB MEASURES ESTIMATED BY UCON, PROX AND UFORM

Ability Group	Person Name	KCTB Score r	Relative Score f = r/L	UCON Ability b_1	UCON Error s_1	PROX Ability b_2	PROX Error s_2	UFORM Ability b_3	UFORM Error s_3	Differences PROX-UCON $b_2 - b_1$	Differences UFORM-UCON $b_3 - b_1$
	3M	1	.04	-5.8	1.2	- 6.8	2.2	- 6.1	1.2	-1.0	- 0.3
Preschool	6F	3	.13	-3.9	0.8	- 4.2	1.4	- 4.3	0.8	-0.3	- 0.4
	12M	5	.22	-2.8	0.7	- 2.8	1.1	- 3.2	0.7	-0.0	- 0.4
	29M	10	.43	-0.9	0.6	- 0.6	0.9	- 0.8	0.7	0.3	0.1
Primary	35F	11	.48	-0.5	0.6	- 0.2	0.9	- 0.2	0.7	0.3	0.3
	69M	15	.65	1.4	0.7	1.4	1.0	1.7	0.7	0.0	0.3
	88M	18	.78	3.0	0.8	2.8	1.1	3.2	0.7	0.2	0.2
Adult	98F	20	.87	4.3	0.9	4.2	1.4	4.3	0.8	0.1	0.0
	101F	21	.91	5.2	1.0	5.1	1.6	5.0	0.9	0.1	0.2

UCON

See Table 7.3.4

PROX

height h = 0
expansion X = 2.2
length L = 23

$b_2 = 2.2 \ln [r/(23-r)]$

$s_2 = 2.2 [23/r(23-r)]^{1/2}$

UFORM

height h = 0
width w = 11
length L = 23

$b_3 = x_{fw}$ See Table 7.3.1

$s_3 = C^{1/2}_{fw}/23^{1/2}$ See Table 7.3.2

His UFORM ability and error are calculated from his relative score f = r/L = 10/23 = .43 and the values for x_{fw} and $C_{fw}^{1/2}$ found in Tables A and B of the appendix with h = 0, w = 11 and L = 23. Thus

$$b = h + x_{fw}, \qquad x_{fw} = -0.8 \quad \text{from Table A}$$

$$= 0 - 0.8$$

$$= -0.8 \ .$$

and

$$s = C_{fw}^{1/2} / L^{1/2}, \qquad C_{fw}^{1/2} = 3.3 \quad \text{from Table B}$$

$$= 3.3/23^{1/2}$$

$$= 0.7 \ .$$

The last columns of Table 7.3.5 give the difference between PROX or UFORM and UCON. With the exception of the PROX measure for Person 3M, no difference is larger than 0.4 logits. All differences are less than half of the standard errors associated with their ability measures.

Confidence in the use of the UFORM Tables 7.3.1 and 7.3.2 or Appendix Tables A and B depends on a knowledge of their functioning over a variety of typical test situations. Wright and Douglas (1975a) investigated their functioning with a simulation study designed to check on the major threats to the success of these tables in providing useful measures.

The results of their study are summarized by the bounds given in Table 7.3.6 for the extent to which a test can depart in practice from a uniform spacing of item difficulties before measurements based on the assumption of a uniform test become unacceptable. Table 7.3.6 gives the combinations of H - β, W, and L within which the bias in estimating β caused by non-uniformity in item difficulty is less than 0.1 logits.

The amount of leeway shown in Table 7.3.6 may seem surprising, since it allows a random item difficulty of, say, d = 2.0 when uniformity calls for δ = 1.0. But, when h and w are calculated from a test's actual d_i, it is demonstrable that a broad spectrum of test designs is exceptionally robust with respect to random departures from uniformity in item difficulty.

Table 7.3.6 shows that as test length increases beyond 30 items, no reasonable testing situation risks measurement bias large enough to matter. Tests in the neighborhood of 30 items, of width less than 8 logits and which come within 1 logit of their target β are, for all practical purposes, free from bias caused by random deviations in the uniformity of item calibrations of magnitude less than 1 logit. Only when tests are as short as 10 items, wider than 8 logits and more than 2 logits off-target does the measurement bias caused by random non-uniformity of item difficulty exceed 0.2 logits. This means that UFORM measurement tables, even though they are based on the assumption of perfectly uniform tests, can be used to transform scores into measures in most practical situations.

TABLE 7.3.6

PERFORMANCE OF UFORM PROCEDURE FOR TESTS LESS THAN 8 LOGITS WIDE

Maximum Item Bias	Maximum Off-Target	Minimum Test Length	Maximum Measurement Bias	
$\lvert d_i - \delta_i \rvert$	$\lvert H - \beta \rvert$	L	BIAS	BIAS/SEM
1.0	2	10	.2	.4
	1	30	.1	.3
0.5	2	10	.1	.2
	1	30	.1	.2

BIAS = The average measurement bias in 100 replications of a test in which the random departures from a uniform distribution of item difficulties are bounded by $\lvert d_i - \delta_i \rvert$.

7.4 INDIVIDUALIZED TESTING

The need for individualized testing becomes obvious whenever we encounter a situation in which inappropriate items have been given to a person. The solution to this problem is to tailor tests to persons. The construction of a bank of calibrated items makes the efficient implementation of tailored testing simple. The uniformity of measurement precision near the center of tests of typical height and width shows that we need only bring the selected items to within a logit of their intended target to achieve "good enough" tailoring. This can be done in various ways.

Status Tailoring. Information about grade placement or age will often be sufficient to tailor a school test. Prior knowledge of the approximate grade placement of the target group or pupil and of the variable's grade norms can be used to determine an appropriate segment of items. Normative data in a variety of school subjects suggests that typical within grade standard deviations are about one logit. When this is so, even a rough idea as to a pupil's within grade quartile provides more than enough information to design a best test for that pupil.

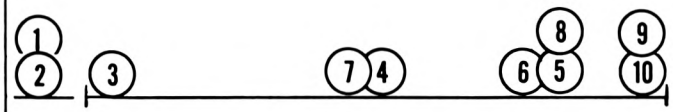
Performance Tailoring. Where grade or age information are not sufficient, tailoring can be accomplished with a pilot test of 5 to 10 items spread out enough in difficulty to cover the widest expected target. If the pilot test were set up to be self-scoring, then pupils could use their number right to guide themselves into a second test specifically tailored to the ability level implied by their pilot test score.

Self-Tailoring. A third even more individualized scheme may prove practical in many circumstances. The person to be measured is given a booklet of items presented in order of uniformly increasing difficulty and asked to find their own best working level. Testing begins when the person finds items hard enough to interest them but easy enough to master. Testing continues into more difficult items until the person decides that the level of difficulty is beyond their ability. The self-tailored test on which this person is then measured is the continuous segment of items attempted.

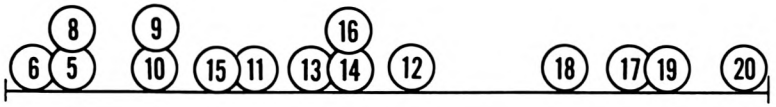
FIGURE 7.5.1

ITEM DISTRIBUTION FOR THREE SEQUENTIAL FORMS MEASURING THE KCT VARIABLE

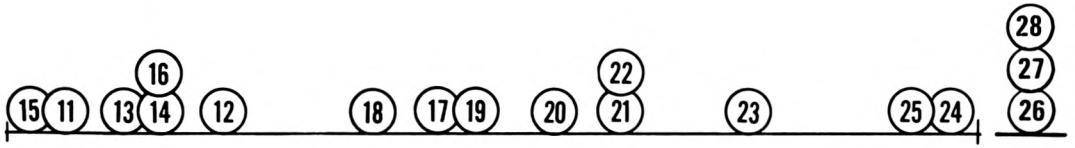
1. Preschool Form, Items 1 - 10



2. Primary Form, Items 5, 6, 8-20



3. Adult Form, Items 11 - 28





This approach is self-adapting to individual variations in speed, test comfort and level of productive challenge. The large variety of different test segments which can result are easy to handle. The sequence number of the easiest and hardest items attempted and the number of correct responses between them can be read off a self-scoring answer form and converted into a measure and its standard error merely by looking up these three statistics in a simple one-page table made to fit with the booklet of items used in testing.

Self-tailored testing corresponds to the use of basal and ceiling levels on individually administered tests like the Stanford-Binet. The only difference is that, with the self-tailored test, the segment of items administered is determined by the person taking the test rather than by an examiner.

7.5 STATUS TAILORING

To illustrate status tailoring we allocated our KCTB items to three sequential forms. The Preschool Form is aimed at preschool children. The Primary Form is aimed at primary school children. The Adult Form is for persons beyond primary school. Figure 7.5.1 shows the distribution of items into these three forms.

The Preschool Form is composed of the first 10 items. Only Items 3 through 10 are calibrated because virtually everyone tested so far has gotten Items 1 and 2 correct. The Primary Form is composed of Items 5, 6 and 8 through 20 to cover the middle range of the variable. The Adult Form is composed of Items 11 through 25, the hardest items calibrated, and Items 26, 27 and 28 which are so hard that no one tested so far has gotten them correct.

Notice that we can include these five "out-of-bound" items in our test forms without impairing our measurements in any way. This is because we can focus our measurements on the portion of the test which is both taken by the person and made up of calibrated items while letting extreme items continue to work for us as the conceptual boundaries of the KCT variable. If eventually we encounter persons who fail Items 1 or 2 or who pass Items 26, 27 or 28, then we will also be able to calibrate these items onto the KCT variable and use responses to them in our measurements.

The items for each of the three forms and their corresponding item difficulties, where known, are given in Table 7.5.1. Below the items in each form are that form's test characteristics: height h, width w and length L.

These three test forms were applied to the nine persons. Table 7.5.2 shows how each person scored on each of the forms. Persons 3M and 6F could be measured on only the Preschool Form while persons 98F and 101F could be measured on only the Adult Form. Person 12M produced a measurable record on the Preschool and Primary Forms. Persons 69M and 88M produced measurable records on the Primary and Adult Forms. Persons 29M and 35F produced measurable records on all three forms.

For Person 29M with relative score .75 on the Preschool Form (h = -3.3, w = 4, L = 8) we look up $x_{fw} = 1.4$ and $C_{fw}^{1/2} = 2.6$ in Tables A and B to find the estimate

$$b = -3.3 + 1.4 = -1.9$$

with standard error

$$s = 2.6/8^{1/2} = 0.9 \quad .$$

TABLE 7.5.1

TEST STATISTICS OF THREE SEQUENTIAL FORMS MEASURING THE KCT VARIABLE

PRESCHOOL FORM		PRIMARY FORM		ADULT FORM	
Item Name	Item Difficulty	Item Name	Item Difficulty	Item Name	Item Difficulty
1	*				
2	*				
3	- 6.2				
4	- 4.1				
5	- 2.6	5	- 2.6		
6	- 2.7	6	- 2.7		
7	- 4.3				
8	- 2.6	8	- 2.6		
9	- 2.1	9	- 2.1		
10	- 2.1	10	- 2.1		
		11	- 1.0	11	- 1.0
		12	- 0.1	12	- 0.1
		13	- 0.9	13	- 0.9
		14	- 0.5	14	- 0.5
		15	- 1.5	15	- 1.5
		16	- 0.8	16	- 0.8
		17	1.9	17	1.9
		18	1.4	18	1.4
		19	2.0	19	2.0
		20	2.9	20	2.9
				21	3.3
				22	3.3
				23	4.5
				24	6.3
				25	5.8
				26	**
				27	**
				28	**

Test Characteristics of the Calibrated Items

	Preschool Form	Primary Form	Adult Form
Height:	h = −3.3	h = −0.6	h = 1.8
Width:	w = 4	w = 6	w = 8
Length:	L = 8*	L = 15	L = 15**

* Items 1 and 2 were too easy to calibrate

** Items 26, 27 and 28 were too hard to calibrate

Items 1, 2, 26, 27 and 28 cannot be used for measurement because their difficulty levels have so far eluded calibration.

TABLE 7.5.2

MEASURING THE THREE ABILITY GROUPS WITH THE THREE SEQUENTIAL FORMS

		Preschool Form (h=−3.3, w=4, L=8)				Primary From (h=−.6, w=6, L=15)				Adult Form (h=1.8, w=8, L=15)			
Ability Group	Person Name	Score r_1	Relative Score f=r/8	Ability b_1	Error s_1	Score r_2	Relative Score f=r/15	Ability b_2	Error s_2	Score r_3	Relative Score f=r/15	Ability b_3	Error s_3
	3M	1	.13	−5.7	1.1	0	0	*		0	0	*	
Preschool	6F	3	.38	−3.9	0.8	0	0	*		0	0	*	
	12M	5	.63	−2.6	0.8	3	.20	−2.8	0.8	0	0	*	
	29M	6	.75	−1.9	0.9	7	.47	−0.8	0.7	4	.27	−0.2	0.8
Primary	35F	7	.88	−0.8	1.2	8	.53	−0.4	0.7	4	.27	−0.2	0.8
	69M	8	1.00	**		12	.80	1.6	0.8	7	.47	1.6	0.8
	88M	8	1.00	**		14	.93	3.0	1.1	10	.67	3.2	0.8
Adult	98F	8	1.00	**		15	1.00	**		12	.80	3.4	0.8
	101F	8	1.00	**		15	1.00	**		13	.87	5.2	0.9

* Score 0 is too low for measurement

** Scores 8 and 15 are too high for measurement

For Person 29M's relative score of .47 on the Primary Form (h = -0.6, w = 6, L = 15) we look up $x_{fw} = -0.2$ and $C_{fw}^{1/2} = 2.6$ to find the estimate

$$b = -0.6 - 0.2 = -0.8$$

with standard error

$$s = 2.6/15^{1/2} = 0.7 \quad .$$

For Person 29M's relative score .27 on the Adult Form (h = 1.8, w = 8, L = 15) we look up $x_{fw} = -2.0$ and $C_{fw}^{1/2} = 3.0$ to estimate

$$b = 1.8 - 2.0 = -0.2$$

with standard error

$$s = 3.0/15^{1/2} = 0.8 \quad .$$

Even though only one of the forms taken is best focused on a person and so produces their "best" measure, still we can see in Table 7.5.2 that, in spite of the wide variation in score on different forms, the measures for a given person are, for the most part, comparable. Person 29M produces the greatest variation in measures over these three forms. His three relative scores of .75, .47 and .27 vary widely in response to the variation in difficulty of the three forms. According to our model his three measures of -1.9, -0.8 and -0.2 ought to be statistically equivalent, even though they may seem to vary more than we might like. When their variation is evaluated in the light of their standard errors of 0.9, 0.7 and 0.8 we see that the lowest estimate of -1.9 on the Preschool Form plus one of its standard errors and the highest estimate of -0.2 on the Adult Form minus one of its standard errors touch at -1.0.

Table 7.5.3 shows for each person their ability measure on the total KCTB test and their ability measure on each of the three sequential forms. The difference between each test form and the KCTB is given at the right of the table. When these differences are compared to the errors associated with them it can be seen that all of the differences are less than half a standard error except for those of Persons 29M and 98F.

The standard errors for each ability for the KCTB and the three test forms are given in Table 7.5.4. These values are stable and consistent over forms for the nine persons.

7.6 PERFORMANCE TAILORING

To illustrate performance tailoring we developed a Pilot Form of seven items from KCTB. Figure 7.6.1 shows the distribution of these seven items. They were selected to be as uniform as possible over the 15 logit range of the KCT variable. Table 7.6.1 gives their item difficulties and the Pilot Form test characteristics. Height is centered at 0.0. The effective width is 15 logits.

To demonstrate performance tailoring with this Pilot Form we will use the performances of our nine persons on the Pilot Form to indicate the sequential form most appropriate for measuring each of them. Then, we will measure them on the indicated sequential form and compare their "performance tailored" measure with their measure based on all 23 KCTB items.

TABLE 7.5.3

COMPARING MEASURES FROM THE SEQUENTIAL FORMS WITH MEASURES FROM THE KCT BANK

Ability Group	Person Name	KCTB Ability* L=23 $\hat{\beta}$	Sequential Forms** Preschool L=8 b_1	Primary L=15 b_2	Adult L=15 b_3	Difference Between Measurement on the KCTB and the Sequential Forms $b_1 - \hat{\beta}$	$b_2 - \hat{\beta}$	$b_3 - \hat{\beta}$	KCTB Error $s_{\hat{\beta}}$
	3M	−5.8	− 5.7			0.1			1.2
Preschool	6F	−3.9	− 3.9			0.0			0.8
	12M	−2.9	− 2.6	− 2.8		0.3	0.1		0.7
	29M	−0.9	− 1.9	− 0.8	− 0.2	[− 1.0]	0.1	[0.7]	0.6
Primary	35F	−0.5	− 0.8	− 0.4	− 0.2	− 0.3	0.1	0.3	0.6
	69M	1.4		1.6	1.6		0.2	0.2	0.7
	88M	3.0		3.0	3.2		0.0	0.2	0.8
Adult	98F	4.3			3.4			[− 0.9]	0.9
	101F	5.2			5.2			0.0	1.0

* Calculated by UCON

** Calculated by UFORM

TABLE 7.5.4

COMPARING MEASUREMENT PRECISION FROM THE SEQUENTIAL FORMS WITH MEASUREMENT PRECISION FROM THE KCT BANK

Ability Group	Person Name	KCTB L=23 Error $\hat{\sigma}$	Sequential Forms Preschool L=8 Error s_1	Primary L=15 Error s_2	Adult L=15 Error s_3
	3M	1.2	1.1*		
Preschool	6F	0.8	0.8		
	12M	0.7	0.8	0.8	
	29M	0.6	0.9	0.7	0.8
Primary	35F	0.6	1.2	0.7	0.8
	69M	0.7		0.8	0.8
	88M	0.8		1.1	0.8
Adult	98F	0.9			0.8 *
	101F	1.0			0.9 *

* Discrepancies from KCTB minimum $\hat{\sigma}$ are due to UFORM approximation

In Table 7.6.2 we give for each person their name and KCTB ability. Next we give their performance on the Pilot Form. This includes their pilot score r, relative score f, ability b_1 and error s_1. Then we show the target regions (from $b_1 - s_1$ to $b_1 + s_1$) implied by each Pilot Form performance. This is followed by the sequential form indicated and the resulting ability measure b_2 based on their performance on the indicated sequential form. Finally, we show the difference $(b_2 - \hat{\beta})$ between the KCTB measure $\hat{\beta}$ and the sequential form measure b_2 together with the KCTB error $s_{\hat{\beta}}$.

For example, Person 29M had a KCTB ability of -0.9. Using the Pilot Form we found an ability of +1.1 with a standard error of 1.5 indicating a target range of -0.4 to 2.6. Since the Adult Form is targeted at 1.8 logits, it is the sequential form indicated for measuring 29M. On this form he obtained a measure of -0.2 logits.

In Table 7.6.2 we see that for seven of the nine persons the difference between their measure on the KCTB and their measure on a performance-tailored best sequential form differs by less than half a standard error. Persons 29 M and 98F, however, show discrepancies between the measures implied by the KCTB and the sequential test form which are of the order of one standard error.

FIGURE 7.6.1

ITEM DISTRIBUTION OF A PILOT FORM FOR LOCATING PERSONS ON THE KCT VARIABLE

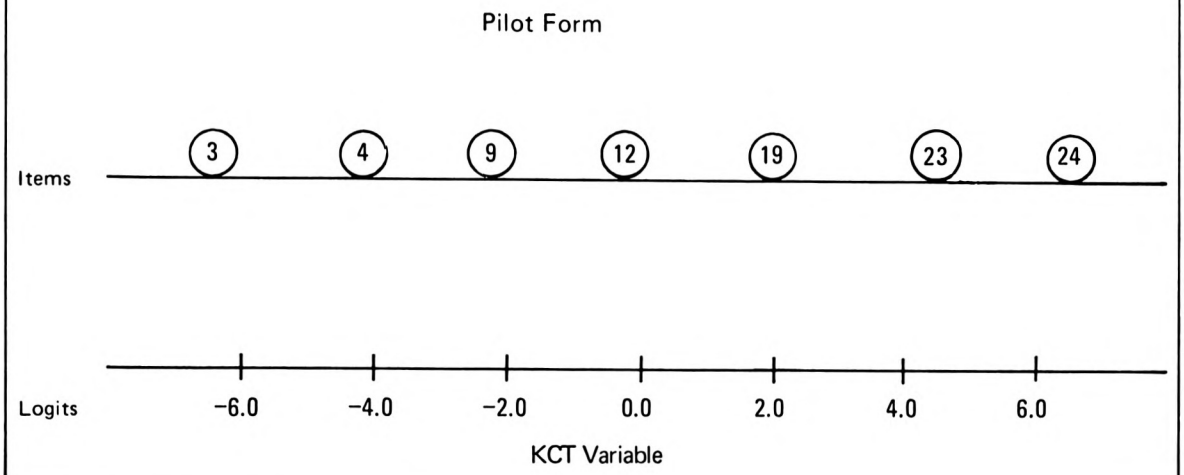


TABLE 7.6.1

TEST STATISTICS OF A PILOT FORM FOR LOCATING PERSONS ON THE KCT VARIABLE

PILOT FORM

Item Name	Item Difficulty
3	−6.2
4	−4.1
9	−2.1
12	−0.1
19	2.0
23	4.5
24	6.3

Height:	h = 0.0
Width:	W = 15
Length:	L = 7

TABLE 7.6.2

LOCATING PERSONS WITH A PILOT FORM FOR MEASUREMENT WITH AN APPROPRIATE SEQUENTIAL FORM

Ability Group	Person Name	KCTB Ability $\hat{\beta}$	Pilot Form: Score r	Pilot Form: Relative Score f=r/7	Pilot Form: Estimated Ability b_1	Pilot Form: Estimated Error s_1	Target Region Implied: Low b_1-s_1	Target Region Implied: High b_1+s_1	Sequential Form Indicated	Measure Obtained from Indicated Sequential Form b_2	Difference Between KCTB and Indicated Sequential Form $b_2-\hat{\beta}$	KCTB Error $s_{\hat{\beta}}$
	3M	− 5.8	1	.14	− 5.5	1.6	− 7.1	− 3.9	Preschool (h=−3.3)	− 5.9	− 0.1	1.2
Preschool	6F	− 3.9	2	.29	− 3.2	1.5	− 4.7	− 1.7	Preschool (h=−3.3)	− 4.0	− 0.1	0.8
	12M	− 2.8	1	.14	− 5.5	1.6	− 7.1	− 3.9	Preschool (h=−3.3)	− 2.5	0.3	0.7
	29M	− 0.9	4	.57	1.1	1.5	− 0.4	2.6	Adult (h=1.8)	− 0.2	0.7	0.6
Primary	35F	− 0.5	4	.57	1.1	1.5	− 0.4	2.6	Adult (h=1.8)	− 0.2	0.3	0.6
	69M	1.4	5	.71	3.2	1.5	1.7	4.7	Adult (h=1.8)	1.6	0.2	0.7
	88M	3.0	5	.71	3.2	1.5	1.7	4.7	Adult (h=1.8)	3.2	0.2	0.8
Adult	98F	4.3	5	.71	3.2	1.5	1.7	4.7	Adult (h=1.8)	3.4	− 0.9	0.9
	101F	5.2	6	.86	5.5	1.5	4.0	7.0	Adult (h=1.8)	5.2	0.0	1.0

7.7 SELF-TAILORING

In order to develop an example of self-tailoring we return to the original person response records of 1's and 0's to see how individualized response patterns might emerge from a matrix such as Table 2.3.1. We want to select a record for each of our nine persons that approximates a self-tailored sequence of items. To obtain these individualized segments we established basal levels, the point at which a particular person might begin taking items, at three successes prior to the first failure. We set ceiling levels, the point at which a person might stop taking items, at three successive failures. This produced a unique self-tailored sequence of items for each of our nine persons.

In Table 7.7.1 we show the response patterns of these self-tailored tests. The first item Person 29M missed, for example, was Item 6. Thereafter he continued passing and failing items until he failed Items 17, 18, and 19 successively. This defined a self-tailored segment for him ranging from Item 3 through Item 19. On these 17 items he had a score of r = 10.

In Table 7.7.2 we compute a measure for each person based upon their self-tailored test segment. In order to do this computation we determine for each self-tailored segment its test characteristics h, w and L. These test characteristics for each person's self-tailored segment are given on the left of Table 7.7.2.

Thus Person 29M has a relative score of 10 on his self-tailored segment of 17 items. Since his segment has a width w = 8 this score of 10 produces a relative ability measure of 0.7 logits which when adjusted for the height of his segment (h = −1.5) yields an ability estimate of −0.8 with a standard error of 0.7. This ability estimate is only 0.1 logits away from his KCTB ability estimate of −0.9 with error 0.6. Inspection of the differences given in Table 7.7.2 between measures on each self-tailored segment and their corresponding KCTB measures shows that all of the measures obtained by self-tailoring are close to the ability measures obtained by the KCTB.

In Table 7.7.3 we show, for each type of tailoring, the efficiency in item usage for each of our nine persons. We see that a considerable number of items can be saved without much diminishing the accuracy of ability estimates.

Person 29M with a 17 item self-tailored segment requires the most items, yet even this segment is 6 items less than the total 23 KCTB items and there is virtually no loss of measurement accuracy. Person 3M produces almost as precise an estimate with only 4 self-tailored items as can be obtained for him by using all 23 of them. This saves 19 items. Person 3M, however, is at the extreme low end of the KCT variable. As a result only the four easiest items are relevant to measure his ability. Were additional easy items available, we could use them to advantage with Person 3M to improve the precision of his measure.

The self-tailored procedure always achieves the most efficient item utilization. This is especially so when making measures at extremes, in this case, beyond ± 4 logits on the KCTB ability scale. However, while appreciating this apparent efficiency, we must also realize that the items saved are items inappropriate for their target. Our real goal is to make measurements sufficiently accurate to be useful. Accuracy depends on the number of items used which are near enough to the person to be measured so that each item makes an adequate contribution to the estimated measure. This means that we want items to be within a logit of their target. Once the items are brought this near their target, all further considerations of accuracy, and hence of efficiency, boil down to the question of how many of these "tailored" items it is practical for the person to attempt.

TABLE 7.7.1

SELF-TAILORED RESPONSE SEQUENCES FROM KCTB

		Item Name and Difficulty (in difficulty order)																						Self-Tailored Segment	
		#3	#7	#4	#6	#5	#8	#9	#10	#15	#11	#13	#16	#14	#12	#18	#17	#19	#20	#21	#22	#23	#25	#24		
Ability Group	Person Name	−6.2	−4.3	−4.1	−2.7	−2.6	−2.6	−2.1	−2.1	−1.5	−1.0	−0.9	−0.8	−0.5	−0.1	1.4	1.9	2.0	2.9	3.3	3.3	4.5	5.8	6.3	Length L	Score r
	3M	1	0	0	0																				4	1
Preschool	6F	1	1	1	0	0	0																		6	3
	12M	0	1	1	1	1	1	0	0	0															9	5
	29M	1	1	1	0	1	1	1	0	1	0	1	1	0	1	0	0	0							17	10
Primary	35F			1	1	1	0	1	1	1	0	1	1	0	1	0	0	0							15	9
	69M							1	1	1	0	1	1	1	1	0	0	1	1	0	0	0			15	9
	88M													1	1	1	0	1	1	0	1	0	0	0	11	6
Adult	98F																		1	1	1	0	0	0	6	3
	101F																			1	1	1	0	0	5	3

Unlisted responses are either all "1" to the left or all "0" to the right

TABLE 7.7.2

MEASUREMENTS FROM SELF–TAILORED RESPONSE SEQUENCES

Ability Group	Person Name	KCTB Ability $\hat{\beta}$	Self-Tailored Segment Test Characteristics: Height h	Width w	Length L	Self-Tailored Measurement: Score r	Relative Score f = r/L	Relative Ability x_{fw}	Error Coefficient $C^{1/2}_{fw}$	Ability $b=h+x_{fw}$	Error $s = C^{1/2}/L^{1/2}$	Difference Between KCTB and Self-Tailoring: Ability Difference $b-\hat{\beta}$	KCTB Error $s_{\hat{\beta}}$
	3M	−5.8	−4.3	4	4	1	.25	−1.4	2.6	−5.7	1.3	0.1	1.2
Preschool	6F	−3.9	−3.8	4	6	3	.50	0.0	2.3	−3.8	0.9	0.1	0.8
	12M	−2.8	−3.1	4	9	5	.56	0.3	2.3	−2.8	0.8	0.0	0.7
	29M	−0.9	−1.5	8	17	10	.59	0.7	2.9	−0.8	0.7	0.1	0.6
Primary	35F	−0.5	−1.0	6	15	9	.60	0.7	2.6	−0.3	0.7	0.2	0.6
	69M	1.9	−0.7	7	15	9	.60	0.7	2.7	1.4	0.7	0.0	0.7
	88M	3.0	2.8	8	11	6	.55	0.4	2.9	3.2	0.9	0.2	0.8
Adult	98F	4.3	4.4	4	6	3	.50	0.0	2.3	4.4	0.9	0.1	0.9
	101F	5.2	4.6	4	5	3	.60	0.6	2.5	5.2	1.1	0.0	1.0

$$h = \sum_{i}^{L} d_i/L, \qquad w = [\,(d_L + d_{L-1} - d_2 - d_1)/2\,]\,[\,L/(L-2)\,] \text{ rounded to nearest integer}$$

TABLE 7.7.3

MEASUREMENT EFFICIENCIES POSSIBLE WITH THREE TYPES OF TAILORED TESTING

Ability Group	Person Name	KCTB Ability $\hat{\beta}$	KCTB L	Items Used: Status Tailoring L_1	Items Used: Performance Tailoring L_2	Items Used: Self-Tailoring L_3	Items Saved: Status Tailoring $L-L_1$	Items Saved: Performance Tailoring $L-L_2$	Items Saved: Self-Tailoring $L-L_3$	Measurement Precision: KCTB $\hat{\sigma}$	Measurement Precision: Status s_1	Measurement Precision: Perf. s_2	Measurement Precision: Self s_3
	3M	−5.8	23	8	12	4	15	11	19	1.2	1.1*	1.1*	1.3
Preschool	6F	−3.9	23	8	12	6	15	11	17	0.8	0.8	0.8	0.9
	12M	−2.8	23	8	12	9	15	11	14	0.7	0.8	0.8	0.8
	29M	−0.9	23	15	18	17	8	5	6	0.7	0.7	0.8*	0.7
Primary	35F	−0.5	23	15	18	15	8	5	8	0.7	0.7	0.8*	0.7
	69M	1.4	23	15	18	15	8	5	8	0.7	0.8	0.8	0.7
	88M	3.0	23	15	18	11	8	5	12	0.7	0.8	0.8	0.9
Adult	98F	4.3	23	15	18	6	8	5	17	0.8	0.8	0.8	0.8
	101F	5.2	23	15	18	5	8	5	18	0.9	0.9	0.9	1.1

* Discrepancies from KCTB minimum are due to UFORM approximation

7.8 PERSON FIT AND QUALITY CONTROL

During test administration it may appear that an examinee has taken the test as planned. Nevertheless, it is always necessary to examine the actual pattern of responses to see if this pattern does in fact correspond to reasonable expectations.

Consider, for example, a test of 10 items administered in increasing order of difficulty. Table 7.8.1 shows five different ways a score of five might be achieved on such a test. The response patterns of Persons A and B, and even C, seem reasonable. Success occurs on the easier items to the left and failure occurs on the harder items to the right. However, the patterns of Persons D and E are quite implausible. How could it happen that Person D got a score of five by succeeding on the five most difficult items while at the same time failing the five easiest items? That is so contradictory to our expectations for a meaningful test record that we cannot take Person D's score of five as the basis for a valid measure of ability. Person D may be smart and careless or dumb and lucky, but one thing is certain, Person D does not have the intermediate ability implied by a score of five.

The response record of Person E also raises questions. If Person E could answer items in the middle range of difficulty correctly including four of the five hardest items, why were the three easiest items missed?

The Rasch measurement model leads to a comprehensive yet easily applied procedure for evaluating the validity of each examinee's record of responses. In this procedure the person's response record is compared with our expectation of what should happen according to the response model. The procedure uses this comparison to calculate a "fit" statistic which indicates the extent to which the person's performance on the test is in accordance with model expectations.

If x_{vi} is the response of person v with tentative measure b_v to item i with bank calibration d_i, and if $x_{vi} = 0$ for an incorrect response or $x_{vi} = 1$ for a correct one, then according to our measurement model

$$z_{vi}^2 = \exp\left[(2x_{vi} - 1)(d_i - b_v)\right] \qquad [7.8.1]$$

is a standard square residual for evaluating the relationship between the observed response x_{vi} and its model expectations given b_v and d_i. According to expectation this z_{vi}^2 should be approximately distributed as chi-square with about $(L - 1)/L$ degrees of freedom where L is the number of items in the test used to estimate b_v. If the set of $\{ z_{vi}^2 \}$ does appear to be distributed this way, then we have no internal reason to invalidate b_v. But if not, we must acknowledge a departure in the data from our expectation and we must see what we can do about it.

Every response x_{vi} in the set of i = 1 to L taken by person v produces its own almost independent z_{vi}^2. We can sum this set of L residuals $\{ z_{vi}^2 \}$ into an approximate chi-square with about $(L - 1)$ degrees of freedom, and for convenience express this chi-square as the standardized statistic

$$t_v = \left[\ell n\,(v_v) + (v_v - 1)\right]\left[(L - 1)/8\right]^{1/2} \sim N(0,1) \qquad [7.8.2]$$

TABLE 7.8.1

FIVE WAYS TO SCORE FIVE ON A TEN ITEM TEST

	Items in order of increasing difficulty										
	Easiest Item									Hardest Item	
Person	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Score
A	1	1	1	1	0	1	0	0	0	0	5
B	1	1	1	0	1	0	1	0	0	0	5
C	1	1	1	0	0	1	1	0	0	0	5
D	0	0	0	0	0	1	1	1	1	1	5
E	0	0	0	1	0	1	1	0	1	1	5

where v_v is the mean square

$$v_v = \sum_i^L z_{vi}^2 /(L-1) \quad . \qquad [7.8.3]$$

The divisor of 8 in Equation 7.8.2 comes from averaging two opposing standardizations of the mean square v. Thus, if

$$t_1 = (v-1)[(L-1)/2]^{1/2} \sim N(0,1)$$

and

$$t_2 = [\ell n\,(v)]\,[(L-1)/2]^{1/2} \sim N(0,1)$$

then

$$t = (t_1 + t_2)/2 = [\ell n\,(v) + (v-1)]\,[(L-1)/8]^{1/2} \sim N(0,1) \quad .$$

In Table 7.8.2 we work out the person fit analysis for the response patterns of Persons 12M, 35F and 88M. Person 12M has a tentative measure of $\hat{b} = -2.8$. For his first item, $d = -6.2$, his response is $x = 0$. These give him a $(d - b)$ difference of

$$(d-b) = [-6.2 - (-2.8)] = -3.4$$

since $(2x-1) = -1$

then $z^2 = \exp[(2x-1)(d-b)] = \exp[-(-3.4)] = \exp(3.4) = 30$

TABLE 7.8.2

CALCULATING PERSON FIT

Person Name	Tentative Ability*	Response Statistic	Person Response Pattern**														
	b	d	−6.2	−4.3	−4.1	−2.7	−2.6	−2.6	−2.1	−2.1	−1.5						
12M	−2.8	x	0	1	1	1	1	1	0	0	0						
		(d−b)	−3.4	−1.5	−1.3	0.1	0.2	0.2	0.7	0.7	1.3						
		z^2	30.0	0.2	0.3	1.1	1.2	1.2	0.5	0.5	0.3						
		d	−4.1	−2.7	−2.6	−2.6	−2.1	−2.1	−1.5	−1.0	−0.9	−0.8	−0.5	−0.1	1.4	1.9	2.0
35F	−0.3	x	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0
		(d−b)	−3.8	−2.4	−2.3	−2.3	−1.8	−1.8	−1.2	−0.7	−0.6	−0.5	−0.2	0.2	1.7	2.2	2.3
		z^2	0.0	0.1	0.1	10.0	0.2	0.2	0.3	2.0	0.5	0.6	1.2	1.2	0.2	0.1	0.1
		d	−0.5	−0.1	1.4	1.9	2.0	2.9	3.3	3.3	4.5	5.8	6.3				
88M	3.2	x	1	1	1	0	1	1	0	1	0	0	0				
		(d−b)	−3.7	−3.3	−1.8	−1.3	−1.2	−0.3	0.1	0.1	1.3	2.6	3.1				
		z^2	0.0	0.0	0.2	3.7	0.3	0.7	0.9	1.1	0.3	0.1	0.0				

* Abilities are the UFORM measures of self-tailored segments listed in column 10 of Table 7.7.2

** Response patterns come from Table 7.7.1

$$z^2 = \exp\,[(2x - 1)(d - b)]$$

TABLE 7.8.3

CALCULATING PERSON FIT: RESIDUAL ANALYSIS

Person Name	Tentative Ability b	Sum of Squares $_z2$	Degrees of Freedom (L−1)	Mean Square v	Fit Statistic t
12M	− 2.8	35.0	8	4.4	4.9*
35F	− 0.3	16.8	14	1.2	0.5
88M	3.2	7.3	10	0.7	−0.7

*Misfit Signal

$$_z2 = \exp[(2x-1)(d-b)]$$

$$v = \Sigma z^2/(L-1)$$

$$t = [\ell n(v) + (v-1)]\ [(L-1)/8]^{1/2}$$

For each other response in Person 12M's tailored segment of 9 items we have given his x, (d − b) and z^2. The residual analysis based upon this row of z^2's for Person 12M leads to

$$\sum_i^9 z_i^2 = 35,$$

which, for 8 degrees of freedom, gives a mean square of v = 4.4 and an approximate normal deviate t = 4.9. The residual analyses for these three persons are summarized in Table 7.8.3.

Notice in Table 7.8.2 that we have used (d − b) rather than the (b − d) used in Chapter 4. This is because the (d − b) form is convenient for the calculation of z^2. Whenever a response is 0, a minus sign is attached to the difference (d − b) which turns it into (b − d). If, however, we keep this sign change in mind, we can use Table 4.3.3 to determine the values in Table 7.8.2. If you use Table 4.3.3, however, you will find that the values in Table 7.8.2 are slightly more exact than the values determined from Table 4.3.3. The difference is greatest on responses which fit well, but these responses play the smallest role in misfit analysis. The sum of squares Σz^2 of 12M based on Table 4.3.3 would be 33 instead of the 35 given in Table 7.8.3. The resulting t would be 4.5 instead of 4.9.

The fit statistic t is distributed more or less normally but with wider tails. In our practical experience the popular rejection level of about two is unnecessarily conservative. The general guidelines we currently use for interpreting t as a signal of misfit are:

If $t < 3$ we accept the measurement of the person as probably valid.

If $3 < t < 5$ we make a careful examination of the response pattern in order to identify and consider possible sources of misfit.

If $t > 5$ we reject the measure as it stands and take whatever steps we can to extract a "corrected" measure from an acceptable segment of the response record, if one exists.

The detailed study of person misfit of course depends on a detailed study of the approximate normal deviates

$$z_{vi} = (2\ x_{vi} - 1)\left\{\exp\left[(2x_{vi} - 1)(d_i - b_v)/2\right]\right\}$$

in the response record in order to track down the possible sources of irregularity.

Since those portions of the $\Sigma\ z^2$ which contribute most to t are the large positive terms, we can streamline the determination of record validity by forming a quick statistic focused on the most surprising responses. Table 4.3.3 (also given as Appendix Table C) shows that the difference between person measure b and item difficulty d must be of the order of $\pm$ 2.0 before z^2 grows larger than 7 or its probability becomes less than 0.12. To reach a probability for a given response of .05 or less we must relax our standard to a (d - b) difference of $\pm$ 3 producing a z^2 of 20.

If we concentrate our attention on surprising responses for which $|d - b| > 3$, then the actual z^2's may be looked up in Table 4.3.3 (or Appendix Table C) and combined with an average value of 1 for all the remaining items in the response segment to produce a crude $\Sigma\ z^2$ for which a crude t can be calculated.

For example, over the 9 responses of Person 12M, there is only one surprise. This is where $(d - b) = -3.4$ and $z^2 = 30$. Combining this value of 30 with eight 1's for the remaining eight items of the test gives us a crude $\Sigma\ z^2 = 38$ and a crude $t = 5.3$. This value for t is not far from the more exact 4.9 we calculated in Table 7.8.2 and leads us to the same conclusion of a significant misfit.

In Table 7.8.4 we summarize the residual analysis for all nine persons. For each person we give the ability measure and standard error from their self-tailored segment of items. Next we give the sum of squares, degrees of freedom, mean square and fit statistic for each person's record. For eight cases we find no evidence of misfit and so we take their measures as plausible. Only the self-tailored segment of Person 12M's record shows a significant misfit. As we saw in Table 7.8.2, this misfit is due entirely to his incorrect response on the first and easiest item in his record. The reason for this incorrect response might be a failure in test taking or a lapse in functioning. In either case we are still interested in the best possible estimate of Person 12M's ability. The problem of extracting the best possible measure from a flawed record will be discussed in Section 7.10.

TABLE 7.8.4

FIT ANALYSIS OF THE NINE PERSONS MEASURED BY SELF-TAILORED TESTS

Group	Person Name	Self-Tailored Ability b	Self-Tailored Error s	Residual Analysis Sum of Squares $\Sigma_z z^2$	Degrees of Freedom (L−1)	Mean Square v	Fit Statistic t
	3M	−5.7	1.3	1.1	3	0.4	−1.0
Preschool	6F	−3.8	0.9	2.4	5	0.5	−1.0
	12M	−2.8	0.8	35.0	8	4.4	4.9 *
	29M	−0.8	0.7	17.6	16	1.1	0.3
Primary	35F	−0.3	0.7	16.8	14	1.2	0.5
	69M	1.4	0.6	20.0	14	1.4	1.0
	88M	3.2	0.9	7.3	10	0.7	−0.6
Adult	98F	4.4	0.9	2.1	5	0.4	−1.1
	101F	5.2	1.1	1.7	4	0.4	−1.0

*Misfit Signal

7.9 DIAGNOSING MISFIT

Consider again the 10 item test with items in order of increasing difficulty imagined for Table 7.8.1. Were we to encounter the pattern produced by Person E, namely

	Score
0 0 0 1 0 1 1 0 1 1	5

we would be puzzled and wonder how this person could answer the hard questions correctly, while getting the first three easiest questions incorrect. Were they "sleeping" on the easy portion of the test?

On the other hand were we to encounter the response pattern

	Score
1 0 1 0 0 0 0 1 1 1	5

our surprise would be as great, but now we might be inclined to explain the irregularity as the result of lucky "guessing" on the three hardest items.

Both the probabilistic nature of the model and our everyday experience with typical response patterns leads us to expect patterns which have a center region of mixed correct and incorrect responses. When we encounter a pattern like

	Score
1 1 1 1 1 0 0 0 0 0	5

it therefore strikes us as "too good to be true." This unexpectedly regular pattern is sometimes produced by persons who work very slowly and carefully refusing to proceed to the next item until they have done everything possible to answer the present item correctly. We will refer to this pattern as "plodding."

Finally, we can also identify a special form of "sleeping" which might better be called "fumbling" in which the incorrect responses are bunched at the beginning of the test suggesting that the person had trouble getting started.

To summarize, we identify the following kinds of response patterns:

		Score
"normal"	1 1 1 1 0 1 0 0 0 0	5
	1 1 1 0 1 0 1 0 0 0	5
"sleeping" or "fumbling"	0 0 0 1 0 1 1 0 1 1	5
"guessing"	1 0 1 0 0 0 0 1 1 1	5
"plodding"	1 1 1 1 1 0 0 0 0 0	5

In Section 7.8 we identified a misfitting response pattern for Person 12M. Now we will investigate misfitting records, such as that of Person 12M, to see how the diagnosis of irregular response patterns might be accomplished. The self-tailored response pattern for Person 12M, with items in order of increasing difficulty, is

	Score
0 1 1 1 1 1 0 0 0	5

The evaluation of this response pattern in Table 7.8.3 shows a significant misfit, $t = 4.9$. In Table 7.9.1 we show the response pattern for Person 12M again and add for each response the probability p of its occurrence under the model. We also give his response pattern in terms of z's in addition to the z^2's. When we plot the z's for Person 12M in Figure 7.9.1 we see what a "sleeping" or "fumbling" response pattern looks like. This figure displays the segment of items responded to. Each item is spaced horizontally along the KCT variable according to its difficulty on the logit scale. Its vertical position is determined by the person's standard residual z produced in response to that item.

The observed response pattern of Person 12M in Figure 7.9.1 shows how the z statistic indicates misfit. Item 3 has a $z = -5.5$ while the other items have z's near their expected value of zero. The effect of Item 3 upon the response pattern of Person 12M can be highlighted by considering the two alternative patterns given in Table 7.9.1 and Figure 7.9.1.

In alternative pattern A we retain a score of five by exchanging the correct response of "1" to Item 8, a relatively hard item, with the incorrect response of "0" to Item 3, the easiest item attempted. Now we have the pattern

	Score
1 1 1 1 1 0 0 0 0	5

TABLE 7.9.1

DIAGNOSING "SLEEPING"

Case Description	Response Statistic	Item Name and Difficulty (in difficulty order)								
		#3	#7	#4	#6	#5	#8	#9	#10	#15
		−6.2	−4.3	−4.1	−2.7	−2.6	−2.6	−2.1	−2.1	−1.5
		Response Pattern								
	x	0	1	1	1	1	1	0	0	0
Person 12M	(2x−1)(d−b)	3.4	−1.5	−1.3	0.1	0.2	0.2	−0.7	−0.7	−1.3
(b = − 2.8)	z^2	30.0	0.2	0.3	1.1	1.2	1.2	0.5	0.5	0.3
	p	.03	.83	.77	.48	.45	.45	.67	.67	.77
	z	−5.5	0.5	0.5	1.1	1.1	1.1	−0.7	−0.7	−0.5
	x	1	1	1	1	1	0	0	0	0
Alternative A	(2x−1)(d−b)	−3.4	−1.5	−1.3	0.1	0.2	−0.2	−0.7	−0.7	−1.3
	z^2	0.0	0.2	0.3	1.1	1.2	0.8	0.5	0.5	0.3
	p	.99	.83	.77	.48	.45	.56	.67	.67	.77
	z	0.2	0.5	0.5	1.1	1.1	−0.9	−0.7	−0.7	−0.5
	x	1	1	1	1	0	0	0	0	1
Alternative B	(2x−1)(d−b)	−3.4	−1.5	−1.3	0.1	−0.2	−0.2	−0.7	−0.7	1.5
	z^2	0.0	0.2	0.3	1.1	0.8	0.8	0.5	0.5	4.5
	p	.99	.83	.77	.48	.56	.56	.67	.67	.18
	z	0.2	0.5	0.5	1.1	−0.9	−0.9	−0.7	−0.7	2.1

$z^2 = \exp[(2x-1)(d-b)]$ $\quad p = 1/(1+z^2)$ $\quad z = (2x-1)\exp[(2x-1)(d-b)/2]$

FIGURE 7.9.1

DIAGNOSING "SLEEPING"

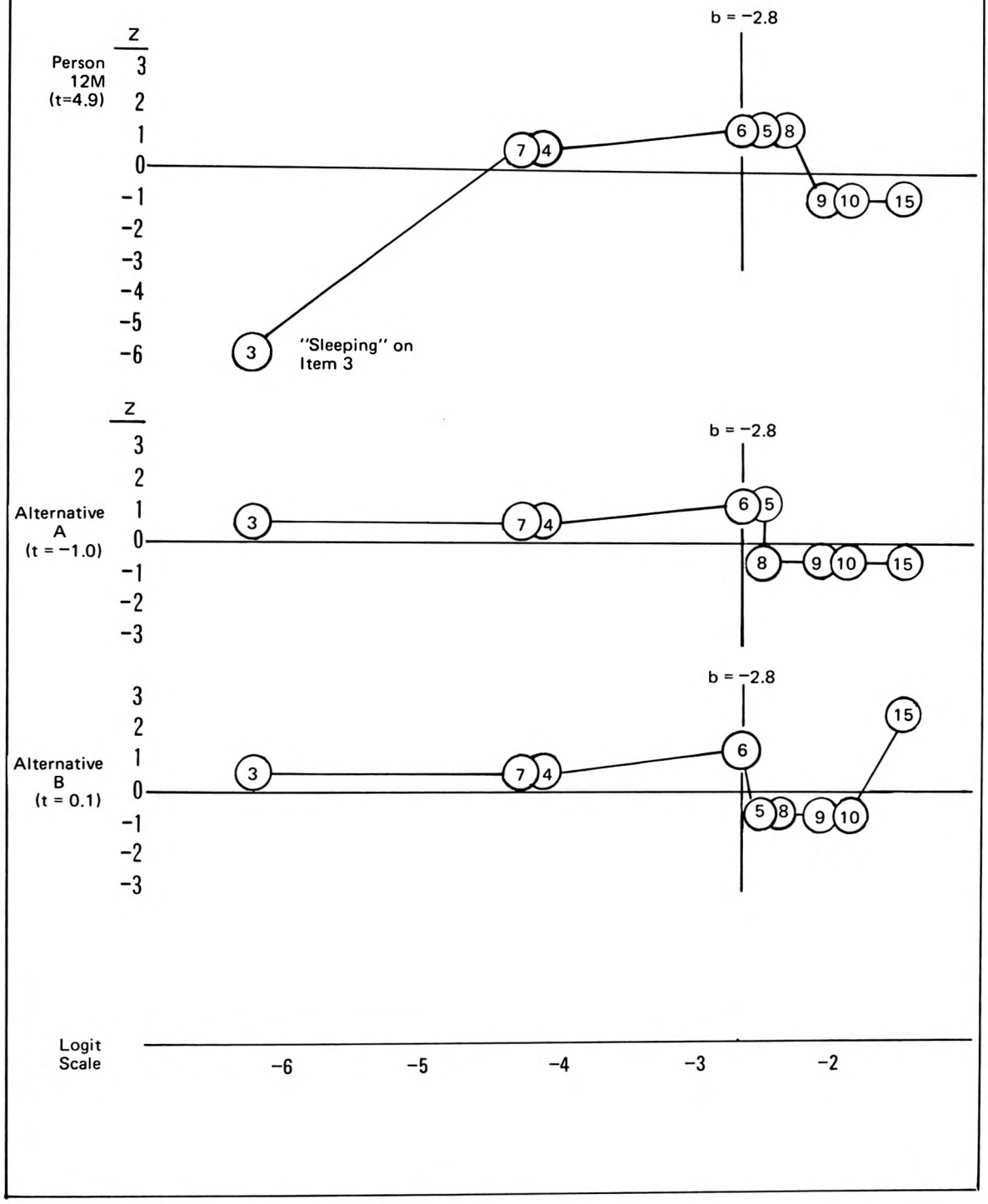


TABLE 7.9.2

RESIDUAL ANALYSIS FOR "SLEEPING" PATTERN OF PERSON 12M

Case Description	Sum of Squares Σz^2	Mean Square v	Fit Statistic t
Person 12M (b = −2.8)	35.0	4.4	4.9 *
Alternative A	4.6	0.6	−1.0
Alternative B	8.5	1.1	0.1

* Misfit Signal

$$z^2 = \exp[(2x - 1)(d - b)]$$

$$v = \Sigma z^2/(L - 1)$$

$$t = [\ell n\,(v) + (v - 1)]\,[L - 1)/8]^{1/2}$$

The misfit statistics for these three patterns are summarized in Table 7.9.2. There we see that Alternative A has a t = −1.0 instead of 12M's t = 4.9.

In Alternative pattern B we exchange the correct response of "1" to Item 5 with the incorrect response of "0" to Item 15, the hardest item in the segment. This produces the alternate response pattern

	Score
1 1 1 1 0 0 0 0 1	5

Interestingly enough, the misfit for the exchange in pattern B is small, only t = 0.1. This is because Item 15, with difficulty d = −1.5, is not as hard in relation to Person 12M's ability of b = −2.8 as Item 3, with difficulty −6.2, is too easy.

In Tables 7.9.3 and 7.9.4 and Figure 7.9.2 we illustrate "sleeping" and "guessing" response patterns using the observed record of Person 88M. To change his response pattern to a sleeping pattern we replace his correct responses to two easy items with incorrect responses and shift these two correct responses to Items 17 and 21, thus keeping the score r = 6. Now we have the response pattern

	Score
0 0 1 1 1 1 1 1 0 0 0	6

which is characteristic of sleeping. This pattern earns t = 9.1 in Table 7.9.4.

TABLE 7.9.3

"SLEEPING" AND "GUESSING" RESPONSE PATTERNS

Case Description	Response Statistic	Item Name and Difficulty (in difficulty order)										
		#14	#12	#18	#17	#19	#20	#21	#22	#23	#25	#24
		−0.5	−0.1	1.4	1.9	2.0	2.9	3.3	3.3	4.5	5.8	6.3
		Response Pattern										
	x	1	1	1	0	1	1	0	1	0	0	0
Person 88M	(2x−1)(d−b)	−3.7	−3.3	−1.8	1.3	−1.2	−0.3	−0.1	0.1	−1.3	−2.6	−3.1
(b = 3.2)	z^2	0.0	0.0	0.2	3.7	0.3	0.7	0.9	1.1	0.3	0.1	0.0
	p	.98	.96	.86	.21	.77	.57	.52	.48	.79	.93	.96
	z	0.2	0.2	0.4	−1.9	0.5	0.9	−1.0	1.1	−0.5	−0.3	−0.2
"Sleeping" Pattern	x	0	0	1	1	1	1	1	1	0	0	0
	(2x−1)(d−b)	3.7	3.3	−1.8	−1.3	−1.2	−0.3	0.1	0.1	−1.3	−2.6	−3.1
	z^2	40.4	27.1	0.2	0.3	0.3	0.7	1.1	1.1	0.3	0.1	0.0
	p	.02	.04	.86	.79	.77	.57	.48	.48	.79	.93	.96
	z	−6.4	−5.2	0.4	0.5	0.5	0.9	1.1	1.1	−0.5	−0.3	−0.2
"Guessing" Pattern	x	1	1	1	1	0	0	0	0	0	1	1
	(2x−1)(d−b)	−3.7	−3.3	−1.8	−1.3	1.2	0.3	−0.1	−0.1	−1.3	2.6	3.1
	z^2	0.0	0.0	0.2	0.3	3.3	1.4	0.9	0.9	0.3	13.5	22.2
	p	.98	.96	.86	.79	.23	.43	.52	.52	.79	.07	.04
	z	0.2	0.2	0.4	0.5	−1.8	−1.2	−1.0	−1.0	−0.5	3.7	4.7

$z^2 = \exp[(2x - 1)(d - b)]$ $\quad p = 1/(1 + z^2)$ $\quad z = (2x - 1)\exp[(2x - 1)(d - b)/2]$

TABLE 7.9.4

RESIDUAL ANALYSIS FOR "SLEEPING" AND "GUESSING" RESPONSE PATTERNS

Case Description	Sum of Squares Σz^2	Mean Square v	Fit Statistic t
Person 88M (b = 3.2)	7.3	0.7	−0.7
"Sleeping" Pattern	71.6	7.2	9.1*
"Guessing" Pattern	43.0	4.3	5.3*

* Misfit Signal

$$z^2 = \exp[(2x - 1)(d - b)]$$

$$v = \Sigma z^2/(L - 1)$$

$$t = [\ell n(v) + (v - 1)][(L - 1)/8]^{1/2}$$

To make a guessing pattern we rearrange responses to form

	Score
1 1 1 1 0 0 0 0 0 1 1	6

for which t = 5.3 in Table 7.9.4. Figure 7.9.2 compares the previously acceptable response pattern of 88M with these alternative unacceptable response patterns characteristic of sleeping and guessing.

A sleeping pattern particularly characteristic of "fumbling" is illustrated in Tables 7.9.5 and 7.9.6 and Figure 7.9.3 where the acceptable response pattern of Person 29M has been altered to show incorrect responses on the first four items of his test, namely Items 3 through 6. While the effect of one incorrect response among these first four items does not produce significant misfit, as seen in the observed pattern for Person 29M, if we make all four items incorrect to illustrate "fumbling" the misfit becomes a significant t = 24.1.

The second pattern illustrated in Figure 7.9.3 is "plodding." In this pattern the person gets every item correct as far as they go and all remaining items incorrect. This can be due to a test-taking style governed by slow and deliberate working habits. While sleeping, guessing and fumbling are indicated by positive values of t, plodding, on the other hand, produces a negative t. The negative value indicates that the observed response pattern fits even better than we expect. It indicates that even the random variability expected by the model is missing!

FIGURE 7.9.2

"SLEEPING" AND "GUESSING" RESPONSE PATTERNS

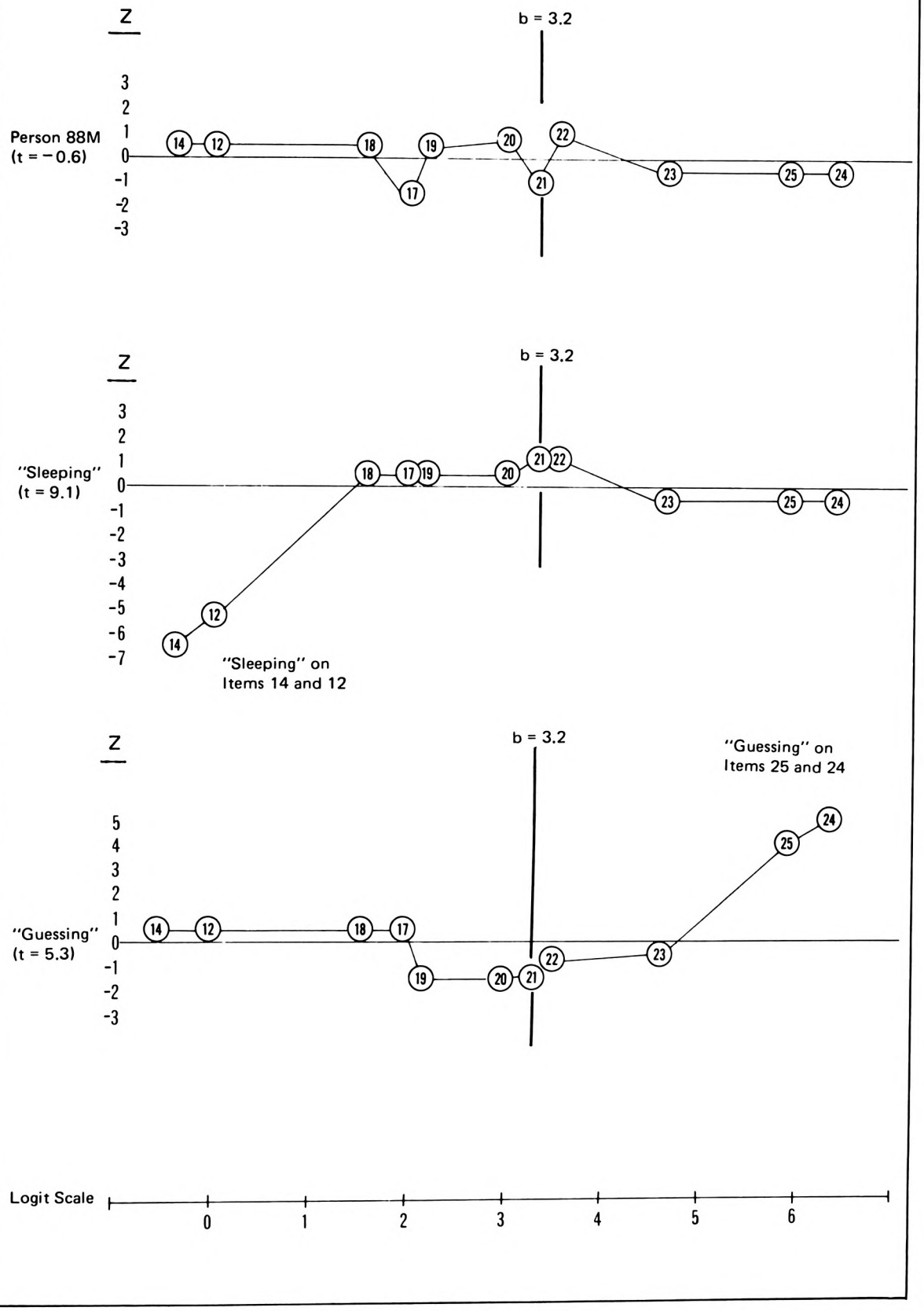


TABLE 7.9.5

"FUMBLING" AND "PLODDING" RESPONSE PATTERNS

		Item Name and Difficulty (in item sequence order)																
		#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19
		−6.2	−4.1	−2.6	−2.7	−4.3	−2.6	−2.1	−2.1	−1.0	−0.1	−0.9	−0.5	−1.5	−0.8	1.9	1.4	2.0
Case Description	Response Statistic	Response Pattern																
	x	1	1	1	0	1	1	1	0	0	1	1	0	1	1	0	0	0
	(2x−1)(d−b)	−5.3	−3.2	−1.7	1.8	−3.4	−1.7	−1.2	1.2	0.1	0.8	0.0	−0.4	−0.6	0.1	−2.8	−2.3	−2.9
Person 29M	z^2	0.0	0.0	0.2	6.0	0.0	0.2	0.3	3.3	1.1	2.2	1.0	0.7	0.6	1.1	0.1	0.1	0.1
(b = −0.9)	p	1.00	.96	.85	.14	.97	.85	.77	.23	.48	.31	.50	.60	.65	.48	.94	.91	.95
	z	0.1	0.2	0.4	−2.5	0.2	0.4	0.6	−1.8	−1.0	1.5	1.0	−0.8	0.7	1.0	−0.3	−0.3	−0.2
"Fumbling"	x	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0
Pattern	(2x−1)(d−b)	5.3	3.2	1.7	1.8	−3.4	−1.7	−1.2	−1.2	−0.1	0.8	0.0	0.4	−0.6	0.1	−2.8	−2.3	−2.9
	z^2	200.3	24.5	5.5	6.0	0.0	0.2	0.3	0.3	0.9	2.2	1.0	1.5	0.6	1.1	0.1	0.1	0.1
	p	.00	.04	.15	.14	.97	.85	.77	.77	.52	.31	.50	.40	.65	.48	.94	.91	.95
	z	−14.2	−5.0	−2.3	−2.5	0.2	0.4	0.6	0.6	1.0	1.5	1.0	1.2	0.7	1.0	−0.3	−0.3	−0.2
"Plodding"	x	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
Pattern	(2x−1)(d−b)	−5.3	−3.2	−1.7	−1.8	−3.4	−1.7	−1.2	−1.2	−0.1	0.8	0.0	−0.4	0.6	−0.1	−2.8	−2.3	−2.9
	z^2	0.0	0.0	0.2	0.2	0.0	0.2	0.3	0.3	0.9	2.2	1.0	0.7	1.8	0.9	0.1	0.1	0.1
	p	1.00	.96	.85	.86	.97	.85	.77	.77	.52	.31	.50	.60	.35	.52	.94	.91	.95
	z	0.1	0.2	0.4	0.4	0.2	0.4	0.6	0.6	1.0	1.5	−1.0	−0.8	−1.4	−1.0	−0.3	−0.3	−0.2

$z^2 = \exp[(2x-1)(d-b)]$ $\quad p = 1/(1+z^2)$ $\quad z = (2x-1)\exp[(2x-1)(d-b)/2]$

TABLE 7.9.6

RESIDUAL ANALYSIS FOR "FUMBLING" AND " PLODDING" RESPONSE PATTERNS

Case Description	Sum of Squares Σz^2	Mean Square v	Fit Statistic t
Person 29M (b = −0.9)	17.6	1.1	0.3
"Fumbling" Pattern	244.7	15.3	24.1*
"Plodding" Pattern	9.0	0.6	−1.3

*Misfit Signal

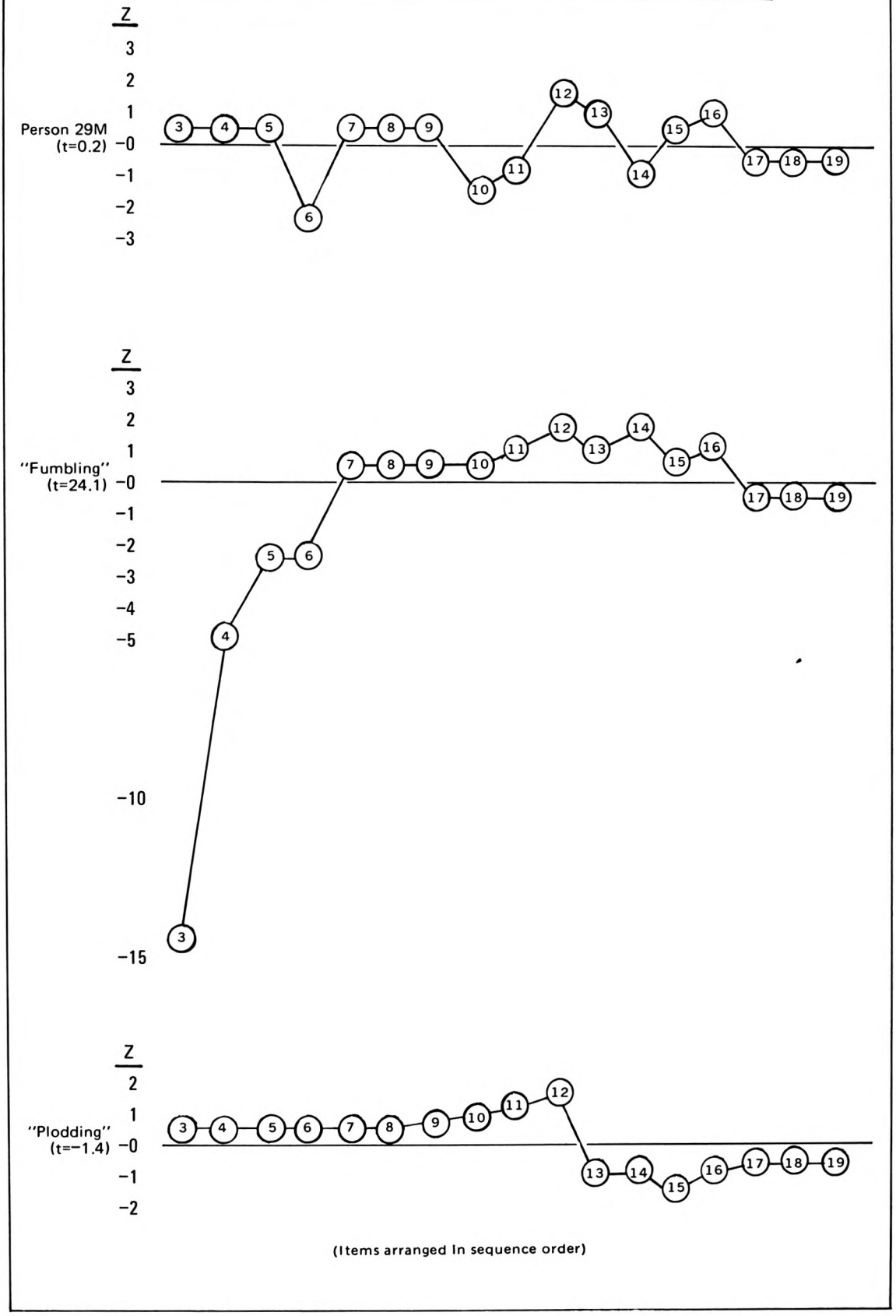
$$z^2 = \exp[(2x - 1)(d - b)]$$

$$v = \sum_{i}^{L} z^2/(L - 1)$$

$$t = [\ell n\ (v) + (v - 1)]\,[(L - 1)/8]^{1/3}$$

FIGURE 7.9.3

"FUMBLING" AND "PLODDING" RESPONSE PATTERNS



7.10 CORRECTING A MEASURE

When we detect significant misfit in a response record, diagnose the response pattern and identify possible reasons for its occurrence, it is finally necessary to decide if an improved measure can or should be determined. Whether such a statistically "corrected" measure is fair for the person or proper in such circumstances cannot be settled by statistics. However, knowing how a measure might be objectively corrected can give us a better understanding of the possible meaning in a person's performance.

We have identified the implausibility of the response of Person 12M to the first item in his test segment given in Tables 7.9.1 and 7.9.2. Were we to decide that this particular response was not typical of Person 12M, we might delete the incorrect response to Item 3 and compute a new ability estimate based on his responses to the remaining eight items. This new calculation of his ability measure is given in Tables 7.10.1 and 7.10.2. The corrected measure $b' = -2.2$ puts Person 12M about 0.6 logits higher on the KCT variable. Figure 7.10.1 shows the effect of this correction on the fit of Person 12M with $t' = -0.8$ instead of $t = 4.9$.

For Person 12M we now have two ability estimates, one at $b = -2.8$ and one at $b' = -2.2$. Which one we decide is the best estimate depends upon how we evaluate the response of Person 12M to Item 3. If we think that this response is implausible and that it is very likely that he would get Item 3 correct, were he to try it again, then we might use the corrected $b' = -2.2$ as his measure. However, if we think, instead, that Person 12M got Item 3 incorrect because of a significant lapse in functioning, then we might consider the $b = -2.8$ as better reflecting his position on the KCT variable. Clinical experience with the KCT variable supports the probability that this lapse is indeed an indicator of impaired functioning and that his incorrect response to Item 3 could be an important element in his evaluation. Consequently, in this case we might well choose the uncorrected measure of $b = -2.8$.

In Tables 7.10.3 and 7.10.4 and Figure 7.10.2 we show the correction of a typical "guessing" pattern. The person's responses to successively more difficult items show four correct responses followed by five incorrect responses and then by two correct ones! This response pattern has a significant misfit of $t = 5.3$. We must ask whether the ability estimate $b = 3.2$ is a good indicator of this person's position on the KCT variable. Given this person's string of five incorrect responses prior to his last two correct ones, we might compute a new estimate with these last two surprising responses removed from the record. With this new truncated pattern $b' = 1.7$ and $t' = -1.2$. Statistical analysis alone cannot tell which estimate is more appropriate, but it can detect and arrange the available information into a concise and objective summary for us to use as part of our evaluation of the person.

Persons who guess may succeed on difficult items more often than their abilities would predict especially on multiple choice items. This makes them appear more able, especially when many items are too difficult for them, because their frequency of success does not decrease as item difficulty increases. A similar but opposite effect occurs when able persons become careless with easy items making these persons appear less able.

Item responses affected by guessing or carelessness actually reflect the simultaneous influence of two variables. There is the ability to be measured, and in addition, there is the tendency to guess or to become careless. The "guessingness" of the item may or may not be a simple function of its difficulty on the main variable or, if a multiple choice

TABLE 7.10.1

CORRECTING THE MEASURE OF PERSON 12M FOR "SLEEPING"

		Item Name and Difficulty (in difficulty order)								
		#3	#7	#4	#6	#5	#8	#9	#10	#15
		−6.2	−4.3	−4.1	−2.7	−2.6	−2.6	−2.1	−2.1	−1.5
Case Description	Response Statistic	Response Pattern								
		delete			b = − 2.8					
Pattern	x	0	1	1	1	1	1	0	0	0
Observed for	(2x−1)(d−b)	3.4	−1.5	−1.3	0.1	0.2	0.2	−0.9	−0.9	−1.5
Person 12M	z^2	30.0	0.2	0.3	1.1	1.2	1.2	0.4	0.4	0.2
	z	−5.5	0.5	0.5	1.1	1.1	1.1	−0.6	−0.6	−0.5
							b′ = − 2.2			
Corrected	x		1	1	1	1	1	0	0	0
Pattern for	(2x−1)(d−b′)		−2.1	−1.9	−0.5	−0.4	−0.4	−0.1	−0.1	−0.7
Person 12M	z^2		0.1	0.2	0.6	0.7	0.7	0.9	0.9	0.5
	z		0.4	0.4	0.8	0.8	0.8	−1.0	−1.0	−0.7

"Sleeping" correction rule: delete $d < (b - 2)$ i.e. $d < -2.8 - 2 = -4.8$

TABLE 7.10.2

RESIDUAL ANALYSIS OF A CORRECTED "SLEEPING" PATTERN

Case	Test Design			Measurement						Residual Analysis		
	Height h	Width w	Length L	Score r	Relative Score $f=r/L$	Relative Ability x_{fw}	Error Coefficient $C_{fw}^{1/2}$	Ability b	Error s	Sum of Squares Σz^2	Mean Squares v	Fit Statistic t
Person 12M	–3.1	4	9	5	.56	0.3	2.3	– 2.8	0.8	35.0	4.4	4.9 *
Corrected Pattern	– 2.8	3	8	5	.63	0.6	2.2	– 2.2	0.8	4.6	0.7	– 0.8

*Misfit Signal

Correction in measure: $(-2.2)-(-2.8) = 0.6$

$$v = \Sigma\, z^2/(L-1) \qquad t = [\ell n(v) + (v-1)]\,[(L-1)/8]^{1/2}$$

FIGURE 7.10.1

CORRECTING THE MEASURE OF PERSON 12M FOR "SLEEPING"

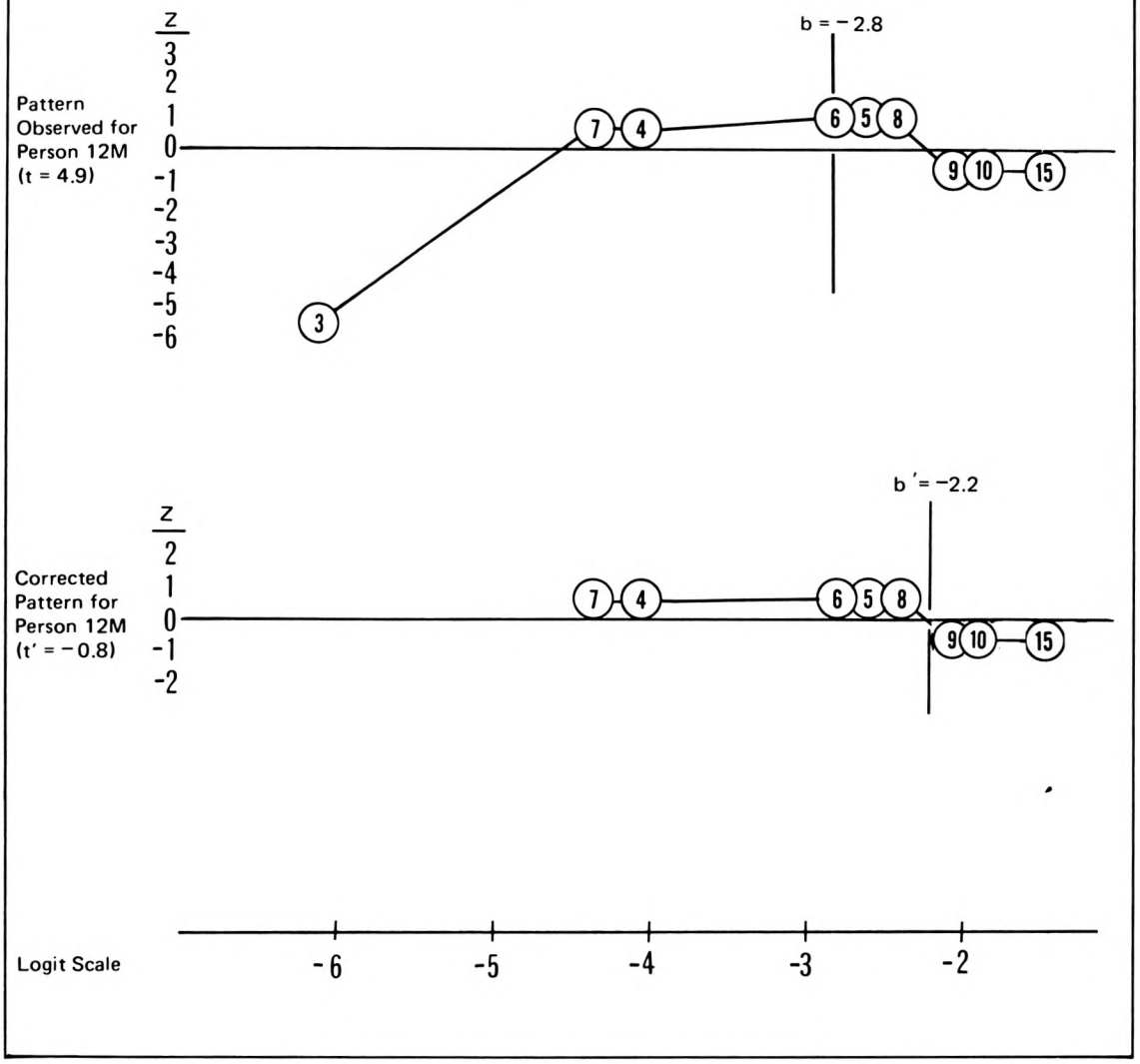


TABLE 7.10.3

CORRECTING A "GUESSING" PATTERN

Case Description	Response Statistic	Item Name and Difficulty (in difficulty order)										
		#14	#12	#18	#17	#19	#20	#21	#22	#23	#25	#24
		−0.5	−0.1	1.4	1.9	2.0	2.9	3.3	3.3	4.5	5.8	6.3
		Response Pattern										
							b = 3.2				delete	
"Guessing"	x	1	1	1	1	0	0	0	0	0	1	1
Pattern	(2x−1)(d−b)	−3.7	−3.3	−1.8	−1.3	1.2	0.3	−0.1	−0.1	−1.3	2.6	3.1
(b = 3.2)	z^2	0.0	0.0	0.2	0.3	3.3	1.4	0.9	0.9	0.3	13.5	22.2
	z	0.2	0.2	0.4	0.5	−1.8	−1.2	−1.0	−1.0	−0.5	3.7	4.7
					b′ = 1.7							
Corrected	x	1	1	1	1	0	0	0	0	0		
Pattern	(2x−1)(d−b′)	−2.2	−1.8	−0.3	−0.2	−0.3	−1.2	−1.6	−1.6	−2.8		
(b′ = 1.7)	z^2	0.1	0.2	0.7	1.2	0.7	0.3	0.2	0.2	0.1		
	z	0.3	0.4	0.9	1.1	−0.9	−0.5	−0.4	−0.4	−0.2		

"Guessing" correction rule: for m-choice items delete $d > [b + \ell n(m - 1)]$,

i.e. if m = 5, then delete $d > 3.2 + \ell n\,(4) = 3.2 + 1.4 = 4.6$

TABLE 7.10.4

RESIDUAL ANALYSIS OF A CORRECTED "GUESSING" PATTERN

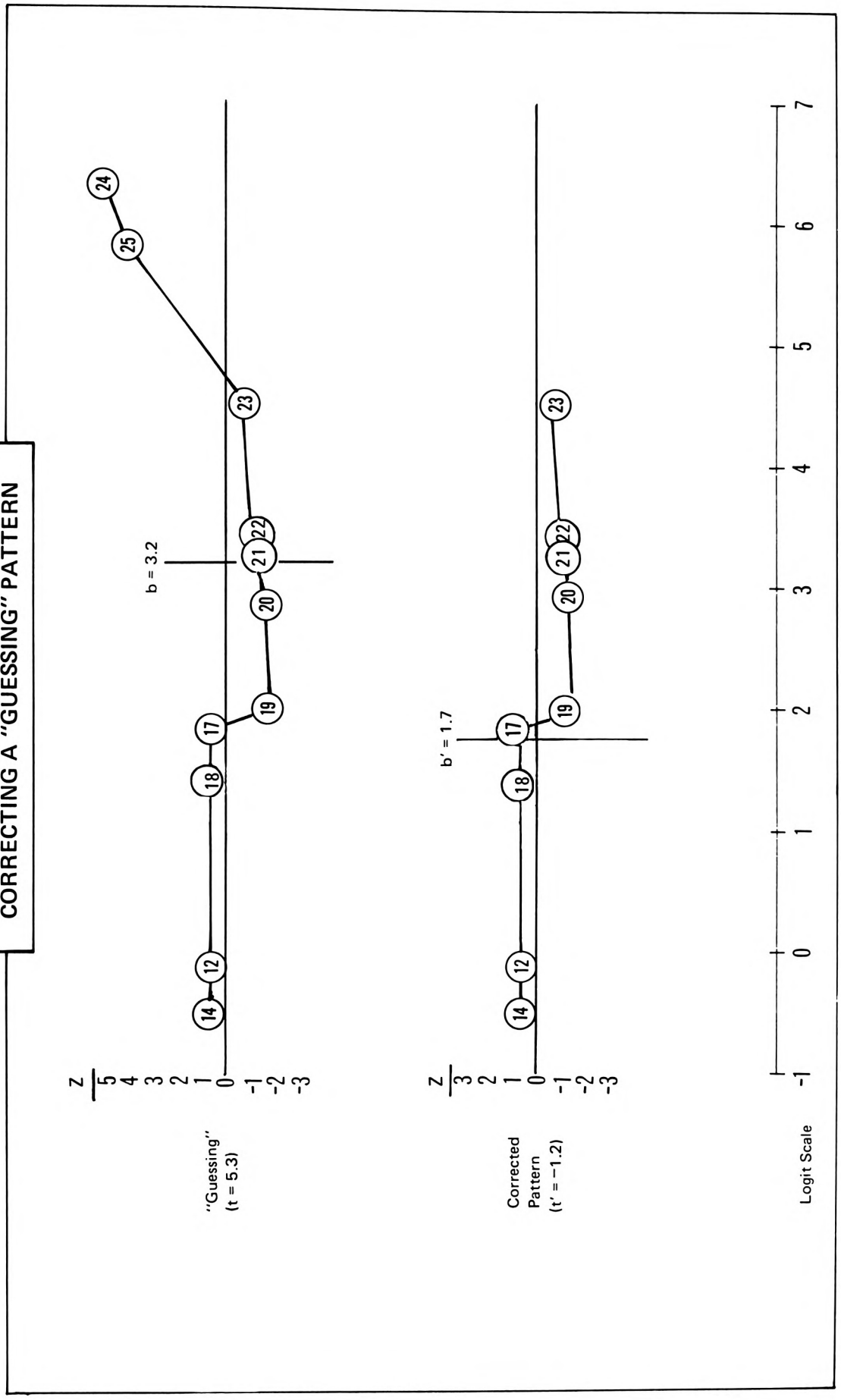
Case Description	Test Design			Measurement						Residual Analysis		
	Height h	Width w	Length L	Score r	Relative Score f=r/L	Relative Ability x_{fw}	Error Coefficient $C_{fw}^{½}$	Ability b	Error s	Sum of Squares Σz^2	Mean Square v	Fit Statistic t
"Guessing" Pattern	2.8	7	11	6	.55	0.4	2.9	3.2	0.9	43.0	4.3	5.3*
Corrected Pattern	2.1	5	9	4	.44	−0.4	2.4	1.7	0.8	3.7	0.5	−1.2

* Misfit Signal

Correction in measure: $1.7 - 3.2 = -1.5$

$$v = \Sigma z^2/(L-1) \qquad t = [\ell n(v) + (v-1)]\,[(L-1)/8]^{½}$$

FIGURE 7.10.2
CORRECTING A "GUESSING" PATTERN



item, of its distractors. For the person being measured, however, two quite different variables are involved. One is their ability, the other is their inclination to guess or their carelessness. The measurement of either variable is threatened by the presence of the other.

In situations where we think that guessing may be influenced by test format as, for example, when we think a person may guess at random over m multiple-choice alternatives, we could use the guessing probability of 1/m as a threshold below which we suppose guessing to occur. To guard our measures against this kind of guessing we can then delete all items from a response record which have difficulty greater than $b + \ell n\,(m - 1)$ where b is the person's initial estimated ability. After these deletions we reestimate the person's ability from the remaining items attempted. If we do this, we are taking the position that when items are so difficult that a person can do better by guessing than by trying, then such items should not be used to estimate the person's ability.

In Tables 7.10.5 and 7.10.6 we show a "fumbling" pattern and its correction. Here we have an increasingly difficult segment of 17 items and a response pattern beginning with four incorrect responses followed by ten correct responses and then three incorrect responses. The pattern seems implausible and significant misfit is identified in $t = 23.1$. Some extraneous factor seems to be influencing the first four responses. It could be a problem of test administration procedures, or of the examinee's test behavior. A corrected response pattern could be formed by deleting the first four incorrect responses and considering only the continuous segment of correct responses and the three incorrect responses which follow them.

The corrected responses resulting from this change show a "plodding" pattern with $t = -3.0$. This pattern produces a considerably higher ability $b' = 1.1$ than the original $b = -0.9$. No final decision can be made on this problem, however, until sufficient clinical or behavioral information is gathered to clarify the meaning of those first four unexpected incorrect responses.

To summarize the statistical aspects of our correction strategy:

a. When the majority of unexpected responses are "incorrect" and $t > 3$ then delete all the "too easy" items $d_i < (b_v - 2)$

   1. Compute the new ability estimate after the deletion of these "too easy" items.
   2. Make another analysis of fit.

b. When the majority of unexpected responses are "correct" and $t > 3$ then delete all the "too hard" items $d_i > [b_v + \ell n\,(m - 1)]$ where m is the number of alternatives.
   1. Compute the new ability estimate after the deletion of the "too hard" items.
   2. Make another analysis of fit.

TABLE 7.10.5

CORRECTING A "FUMBLING" PATTERN

Case Description	Response Statistic	Item Name and Difficulty (in item sequence order)																
		#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19
		−6.2	−4.1	−2.6	−2.7	−4.3	−2.6	−2.1	−2.1	−1.0	−0.1	−0.9	−0.5	−1.5	−0.8	1.9	1.4	2.0
		Response Pattern																
		delete																
"Fumbling"	x	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0
Pattern	(2x−1)(d−b)	5.3	3.2	1.7	1.8	−3.4	−1.7	−1.2	−1.2	−0.1	0.8	0.0	0.4	−0.6	0.1	−2.8	−2.3	−2.9
(b = −0.9)	z^2	200.3	24.5	5.5	6.0	0.0	0.2	0.3	0.3	0.9	2.2	1.0	1.5	0.6	1.1	0.1	0.1	0.1
	z	−14.2	−5.0	−2.3	−2.5	0.2	0.4	0.6	0.6	1.0	1.5	1.0	1.2	0.7	1.0	−0.3	−0.3	−0.2
Corrected	x					1	1	1	1	1	1	1	1	1	1	0	0	0
Pattern	(2x−1)(d−b′)					−5.4	−3.7	−3.2	−3.2	−2.1	−1.2	−2.0	−1.6	−2.6	−1.9	−0.8	−0.3	−0.9
(b′ = 1.1)	z^2					0.0	0.0	0.0	0.0	0.1	0.3	0.1	0.1	0.1	0.1	0.4	0.7	0.4
	z					0.1	0.2	0.2	0.2	0.3	0.5	0.4	0.4	0.3	0.4	−0.7	−0.9	−0.6

"Fumbling" correction rule: delete any continuous segment of "incorrect" responses to easy items that begins a response pattern when they are followed by a subsequent continuous segment of at least as many "correct" responses to more difficult items.

TABLE 7.10.6

RESIDUAL ANALYSIS OF A CORRECTED "FUMBLING" PATTERN

Case Description	Test Description			Measurement						Residual Analysis		
	Height h	Width w	Length L	Score r	Relative Score f=r/L	Relative Ability x_{fw}	Error Coefficient $C_{fw}^{1/2}$	Ability b	Error s	Sum of Squares Σz^2	Mean Square v	Fit Statistic t
"Fumbling" Pattern	−1.6	8	17	10	.59	0.7	2.9	−0.9	0.7	244.7	15.3	24.1*
Corrected Pattern	− 0.8	6	13	10	.77	1.9	2.8	1.1	0.8	2.4	0.2	− 3.0**

Correction in measure: 1.1 − (−0.9) = 2.0

* Misfit Signal

** Plodding Signal

$$v = \Sigma z^2/(L-1) \qquad t = [\ell n(v) + (v-1)]\,[(L-1)/8]^{1/2}$$

8 CHOOSING A SCALE

8.1 INTRODUCTION

Logits are the units of measurement we have used thus far. These units flow directly from the logistic response model which specifies the estimated probability of a correct response by person v to item i as

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)]$$

where b_v is the estimated ability of person v and d_i is the estimated difficulty of item i. It follows that the odds for a correct response are

$$p_{vi}/(1 - p_{vi}) = \exp(b_v - d_i)$$

from which the natural log odds for a correct response becomes

$$\ell n\,[p_{vi}/(1 - p_{vi})] = (b_v - d_i) \quad .$$

These log odds are called "logits" and so differences among items and persons are initially in logit units.

The choice of a unit is entirely arbitrary, but it is absolutely necessary that some unit be chosen. While it is possible to continue to use the initial logits as the units of measurement, this has two disadvantages. Logits involve both negatives and decimals, numerical characteristics which might make them unnecessarily confusing.

The KCT logit scale, for example, extends from -5.8 to +5.2. At the test lengths presently available, standard errors of measurement can be as low as 0.6 logits. We could add a constant such as 10 to do away with the negatives, but we could not avoid decimals by rounding KCT measures in logits to the nearest integer. That rounding would produce a least noticeable difference of almost two standard errors and so could obliterate differences in measures which might be meaningful. Were we to transform the logit scale by first multiplying each value on the scale by 10 and then adding 100, however, we would have a new scale of measures from 42 to 152 which would convey the same information as the initial logit scale but be free from negatives and decimals.

To create a new scale that is free from the inconvenience of decimals we must multiply the logits by a "spacing" factor large enough so that rounding the new units to the nearest integer does not leave behind any useful information. Once this spacing factor is chosen and the unit of our new scale is determined, we can then add a "location" factor to these new integer units that is large enough so that the lowest possible value that can occur is greater than zero. The new scale is defined by determining these two factors. The multiplicative factor establishes the spacing, or units, of the scale. The additive factor establishes the location, or origin, of the scale.

The choice of an additive factor which locates all possible values above zero is usually easy. The choice of a multiplicative factor, however, is worth further consideration. If we want to work in integer units, then we must arrange matters so that any differences

on our new scale smaller than one integer will be meaningless. This requires us to investigate the size of a least meaningful difference.

In addition to determining a least meaningful difference we may also wish to mark easy to remember points like 50, 100 or 500 on our new scale, either at important substantive criteria along the variable or at the typical location of a normative reference group. It is even possible that we will find it useful to relate our new scale unit directly to the probabilities for success predicted by the response model. We may, for example, want to pinpoint movement through memorable response probabilities like .10, .25, .50, .75 and .90 with regular increments of 5, 10, 20, or 25 along our new scale.

Thus, in addition to removing unnecessary negatives and decimals by adding a constant and establishing a least meaningful unit larger than one, we may also organize our scale around normative, substantive or response probability considerations.

8.2 FORMULAS FOR MAKING NEW SCALES

In order to be explicit about how a new scale is determined, we will express its definition as the linear transformation $y = \alpha + \gamma x$ in which x is the logit scale, y is the new scale, α is the location factor for determining the new scale origin and γ is the spacing factor for determining the new scale unit. We make this transformation linear because we want to preserve the interval characteristics of the logits produced by the Rasch model.

Our new measures B and new calibrations D can be expressed in terms of their logit counterparts b and d as

$$B = \alpha + \gamma b \text{ for persons} \qquad [8.2.1]$$

$$D = \alpha + \gamma d \text{ for items} \qquad [8.2.2]$$

The new standard errors of measurement and calibration are

$$SE(B) = \gamma SE(b) \qquad [8.2.3]$$

$$SE(D) = \gamma SE(d) \qquad [8.2.4]$$

This shows how the nature of the new scale depends on the values for α and γ chosen to define it.

In passing let us appreciate again that person ability and item difficulty mark locations on one common variable. In constructing this variable we necessarily work with the calibrations of the items which define it. However, when we use the variable to measure persons we then work with their measures along the variable defined by these items. What a measure tells about a person is the difficulty level of the items on which that person is likely to succeed half the time. In the same way, what a calibration tells about an item is the ability level of persons who are likely to succeed on that item half the time. Thus, were we not reserving the terms "measure" to refer to the location of persons and "calibration" to refer to the location of items, we could as well speak of item difficulty as the measure of the item and of person ability as the calibration of the person.

8.3 THE LEAST MEASURABLE DIFFERENCE

We want to free our new scale from decimals, but we do not want to obliterate useful information. As a result, we need to determine the least measurable difference LMD on our logit scale so that we can choose a spacing factor γ that brings this logit LMD to

at least one integer on our new scale. The nearest any two persons can be in observed scores, without being the same, is one score apart. This is the least observable difference LOD. We need to transform this LOD into its corresponding LMD in logits of ability.

Logit ability b comes from score r through the response model expectation

$$r = \sum_{i}^{L} \left\{ \exp(b_r - d_i)/[1 + \exp(b_r - d_i)] \right\} \quad .$$

As a result LMD must follow LOD at the rate $\partial b/\partial r$ by which scores of r produce measures of b, that is

$$LMD \simeq \frac{\partial b}{\partial r} \; LOD$$

In order to standardize observations with regard to test length L we will generalize from raw score r to relative score f = r/L. Then with LOD = 1 in score r and f = r/L we have LOD = 1/L in relative score f giving

$$LMD \simeq \frac{\partial b}{\partial f} \; LOD = \frac{\partial b}{\partial f} (1/L) \qquad [8.3.1]$$

The Rasch response model gives us the expected relation between relative score f and estimated response probability p_{fi} of

$$f = \sum_{i}^{L} p_{fi} / L$$

in which $p_{fi} = \exp(b_f - d_i)/[1 + \exp(b_f - d_i)]$.

Thus, the rate at which relative score f produces b is

$$\frac{\partial b}{\partial f} = [\sum_{i}^{L} p_{fi} (1 - p_{fi})/L]^{-1} = C_{fw}$$

which turns out to be the error coefficient C_{fw} discussed in Chapters 6 and 7.

This coefficient is subscripted to test width w as well as relative score f because, as we learned in Chapter 6, the exact values of this coefficient depend not only on the relation between test difficulty level and person ability expressed in relative score f = r/L, but also on the width in difficulty covered by the test. This gives us a least measureable difference of

$$LMD \simeq \frac{\partial b}{\partial f} (1/L) = C_{fw}/L \qquad [8.3.2]$$

The way C_{fw} and hence LMD varies with b is pictured in Figure 8.3.1. As the ability measure b moves away from test center and/or the test operating curve flattens the LMD becomes larger. Fortunately the range of values which C_{fw} will have in practice are limited.

When $1/8 < p_{fi} < 7/8 \qquad i = 1, L$

then $4 < C_{fw} < 9$

and $C_{fw} = 6$

can be used as a convenient single working value for C_{fw} (see Table 6.8.1 for details).

This gives us as a working definition of the least measurable difference

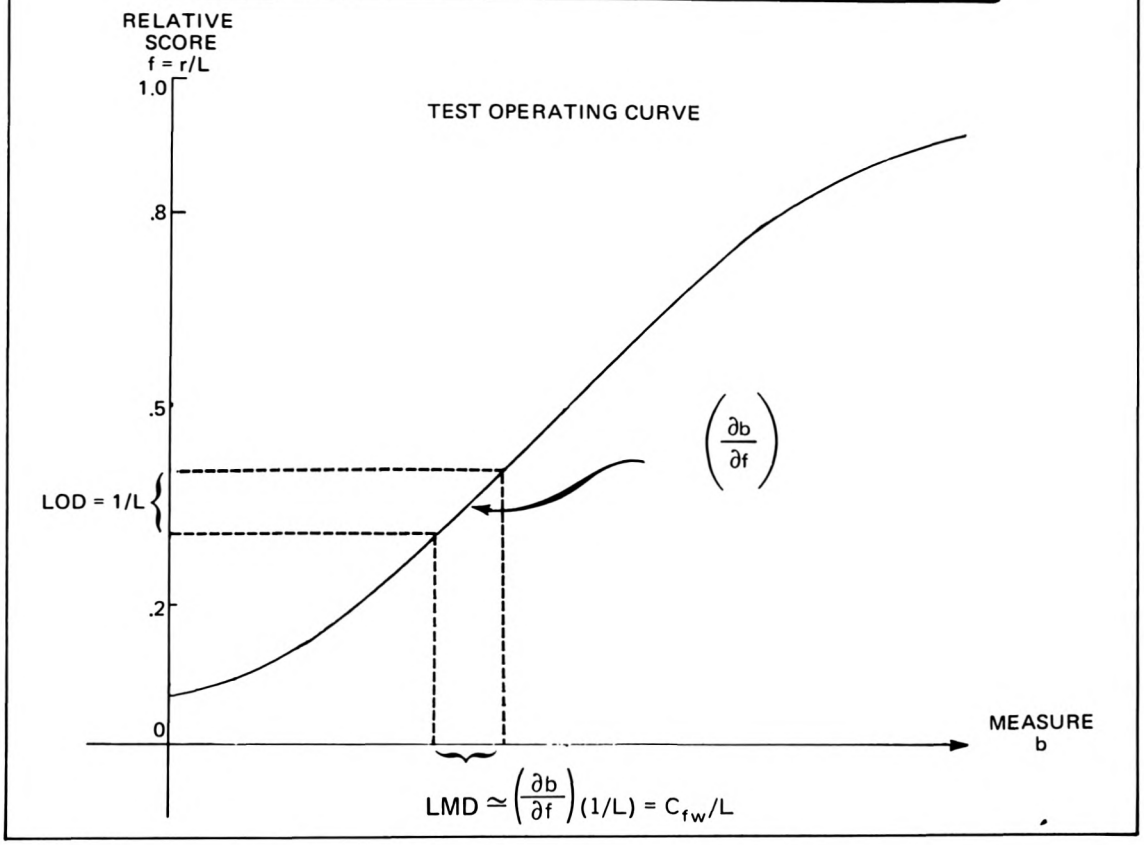
$$LMD = 6/L \qquad [8.3.3]$$

and implies a spacing factor

$$\gamma_{LMD} > L/6 \quad .$$

FIGURE 8.3.1

DETERMINING THE LEAST MEANINGFUL DIFFERENCE



The LMD approximates the smallest possible meaningful unit since it stems from the least observable difference. However, from an estimation point of view, we might consider instead that one standard error of measurement SEM is actually the least "believable" difference. In logits the SEM is related to the LMD as

$$SEM = (LMD)^{1/2} = C_{fw}{}^{1/2}/L^{1/2}$$

which suggest the working value

$$SEM = 2.5/L^{1/2} \qquad [8.3.4]$$

as an alternate basis for determing the spacing factor

$$\gamma_{SEM} > L^{1/2} / 2.5$$

As long as there are more than six items in our test the SEM determines a smaller γ than the LMD since

$$L^{1/2} / 2.5 < L/6 \qquad \text{when} \quad L > 6.$$

An SEM-based scale, which might be simpler numerically, however, will also be somewhat less discriminating in its integer increment than an LMD-based scale. Which choice is preferable in any particular situation cannot be settled by statistical considerations. The choice will inevitably depend on the use to which the measures are to be put.

Finally we might consider the least significant difference between independent measures, whether replications of the same person or comparisons of different persons, as an upper limit on how crude we could allow our new scale to become. To determine this least significant difference LSD we take

$$LSD_{ab} = (SEM_a^2 + SEM_b^2)^{1/2} \simeq (2\,SEM^2)^{1/2}$$

and arrive at the working value

$$LSD = 1.4\,SEM = 3.5/L^{1/2} \qquad [8.3.5]$$

which produces a minimum spacing factor

$$\gamma_{LSD} > L^{1/2} / 3.5$$

As long as the number of items in our test is greater than six, the relative magnitudes of these bases for determining a lower limit for the spacing factor are

$$LMD < SEM < LSD.$$

and so the spacing factors they determine are ordered

$$\gamma_{LMD} > \gamma_{SEM} > \gamma_{LSD}$$

Figure 8.3.2 shows the relationships between LMD, SEM and LSD in logits for the KCTB test of 23 items. The items, their logit values, score equivalents and LOD's at 3 to 4, 12 to 13, 18 to 19 and 20 to 21 along with their corresponding exact LMD's, SEM's and LSD's are shown.

We can compare the exact values in Figure 8.3.2 with the approximations of Equations 8.3.3, 8.3.4 and 8.3.5.

	Minimum γ Implied
$LMD = 6/L = 6/23 = 0.26$	4
$SEM = 2.5/L^{1/2} = 2.5/4.8 = 0.52$	2
$LSD = 3.5/L^{1/2} = 3.5/4.8 = 0.73$	1.5

Because the 13 logit KCTB is unusually wide, these approximations are smaller than the exact values given in Figure 8.3.2. The minimum LMD spacing factor γ indicated by the exact values would be about 2 while the approximations could lead to a minimum γ of 4. Since we would only be in danger of losing information if the approximations led us to a γ of less than 2, we see that even in this extreme situation the approximations do not mislead us.

8.4 DEFINING THE SPACING FACTOR

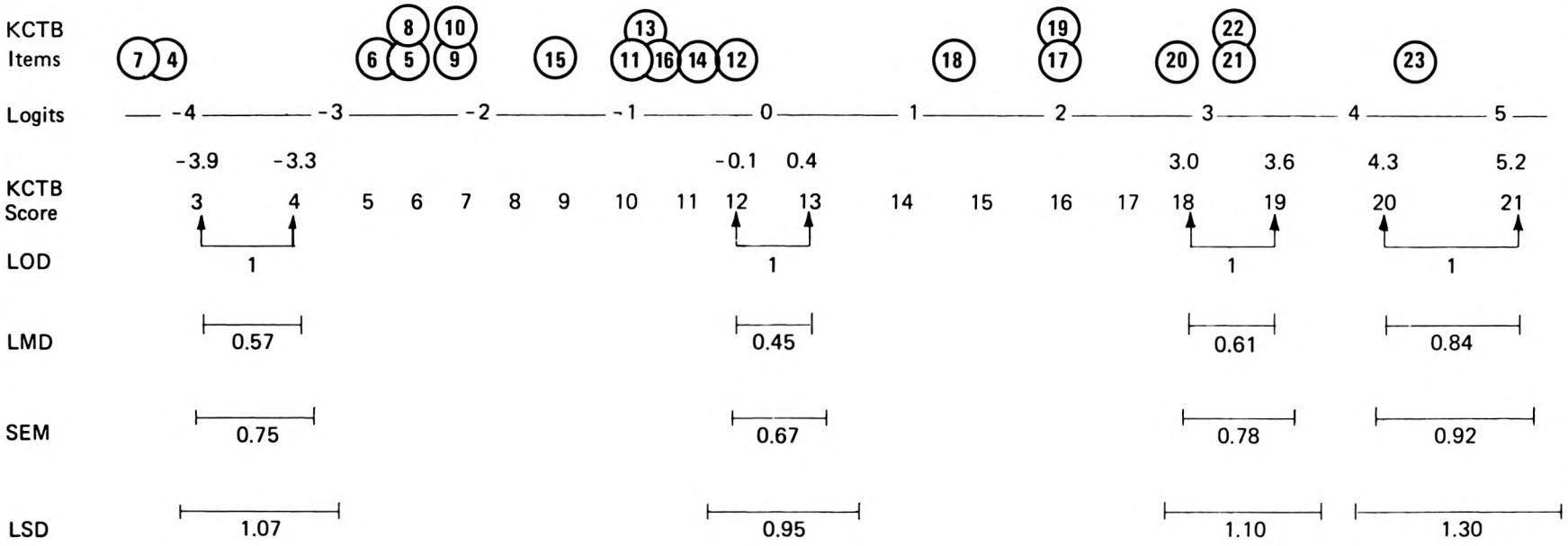
Once we have defined a least meaningful difference in logits, whether it be the least measurable difference LMD (b) = 6/L to maintain maximum observability or the standard error of measurement SEM (b) = $2.5/L^{1/2}$ and its least significant difference LSD (b) = $3.5/L^{1/2}$ to maintain statistical reliability, we can use this least meaningful difference to establish a spacing factor which will make all interpretable differences on our new scale greater than one.

If our aim is to make the least measurable difference in the new scale LMD (B) > 1, then since γ = LMD (B)/LMD (b), it follows as in Equation 8.3.3 that

$$\gamma_{LMD} > L/6 \qquad [8.4.1]$$

FIGURE 8.3.2

LEAST MEANINGFUL LOGIT DIFFERENCES FOR THE KCTB TEST



The scale used here is defined by 23 items. Item 3 at −6.2, Item 25 at 5.8 and Item 24 at 6.3, however, are located beyond the scope of this figure.

The values above are exact. Their rough approximations from Equations 8.3.3, 8.3.4 and 8.3.5 are LMD = 0.26, SEM = 0.52 and LSD = 0.73.

LOD = (r + 1) − r = 1 $\qquad$ LMD = $b_{r+1} - b_r$ $\qquad$ SEM = $LMD^{½}$ $\qquad$ LSD = 1.4 SEM

is the spacing factor which guarantees that no observable differences will be obliterated by rounding to the nearest integer.

Were we interested instead in keeping the spacing factor γ as small as possible, in order to prevent the presentation of scale differences which are statistically unreliable, we might set γ at 1/SEM(b) or even 1/LSD(b) that is

$$\gamma_{SEM} = L^{1/2}/2.5 \qquad [8.4.2]$$

or

$$\gamma_{LSD} = L^{1/2}/3.5 \qquad [8.4.3]$$

Often, however, there will be other considerations which will lead us to allow γ to become even larger than L/6 in order to reach memorable scale intervals like 5, 10, 20, 25, 50 or 100.

To get a rough idea as to typical useful values of γ, we list in Table 8.4.1 values for the least meaningful differences which go with various test lengths. In Table 8.4.1 we see that we would seldom be satisfied with a spacing factor less than 5 and seldom need one larger than 100. Table 8.4.1 suggests that we could work satisfactorily with

$\gamma = 5$ for short classroom tests of 20 or 30 items,

$\gamma = 10$ for typical unit tests of 50 to 60 items

and $\gamma = 20$ or 25 for longer tests of 120 to 150 items.

Only for tests of unusual length, such as 1,000 item examinations, would we want $\gamma = 100$.

TABLE 8.4.1

THE RELATION BETWEEN SPACING FACTORS AND TEST LENGTH

	Least Meaningful Difference			Approximate Spacing Factor γ to Reach an Integer Scale		
Test Length L	Minimum LMD 6/L	SEM $2.5/L^{1/2}$	Maximum LSD $3.5/L^{1/2}$	Maximum γ_{LMD}	γ_{SEM}	Minimum γ_{LSD}
30	0.20	0.46	0.64	[5]	2	2
60	0.10	0.32	0.45	(10)	3	2
120	0.05	0.23	0.32	20	4	3
150	0.04	0.20	0.29	25	[5]	3
300	0.02	0.14	0.20	50	7	[5]
600	0.01	0.10	0.14	100	(10)	7
1200	0.005	0.07	0.10	200	15	(10)

After we decide on γ we apply it to our person b's and item d's to place them on our new scale as B's and D's. While the relation between LMD and SEM in logits of LMD(b) = [SEM(b)]^2 is easy to remember, their relation in the new scale also involves γ. Since

$$LMD(B) = \gamma LMD(b)$$

$$SEM(B) = \gamma SEM(b)$$

but $$LMD(b) = [SEM(b)]^2$$

it follows that in our new scale

$$LMD(B) = [SEM(B)]^2/\gamma \quad \textbf{[8.4.4]}$$

$$SEM(B) = [\gamma LMD(B)]^{1/2} \quad \textbf{[8.4.5]}$$

For example, suppose in order to rescale the KCTB shown in Figure 8.3.2 we chose $\gamma = 5$. Then although in logits

$$LMD(b) = [SEM(b)]^2$$

in our new scale

$$LMD(B) = [SEM(B)]^2/5$$

and $$SEM(B) = [5LMD(B)]^{1/2}$$

Thus while an SEM(b) of $0.75 = 0.57^{1/2}$ goes with an LMD(b) of 0.57, when $\gamma = 5$ then

$$LMD(B) = 5 \times 0.57 = 2.81$$

but $$SEM(B) = 5 \times 0.75 = 3.75 = (5 \times 2.81)^{1/2}$$

8.5. NORMATIVE SCALING UNITS: NITS

If we want our scale to be based on a normative reference, we can use the observed logit mean m and logit standard deviation s of the elected norming sample as factors in a preliminary transformation $d' = (d - m)/s$ and $b' = (b - m)/s$ which puts the norming group mean at zero and the scale unit at one normative standard deviation.

After this preliminary step, we then choose a spacing factor large enough so that meaningful differences become greater than one and at a value which pegs the normative standard deviation at some easy to remember unit such as 10, 20, 50 or even 100. At the same time we choose the location factor α so that the mean of the norming group is also easy to recall, for example at 50, 100 or 500.

Thus to create a norm based scale of normative units or NITs, we use for persons

$$B = \alpha + \gamma (b - m)/s \quad \textbf{[8.5.1]}$$

and for items

$$D = \alpha + \gamma (d - m)/s$$

We then have on the new NITs scale the norming mean $M = \alpha$ and the norming standard deviation $S = \gamma$.

Using the administration of the KCTB to the 68 persons older than 8 as norming data, we have m = 1.3 and s = 1.9 in logits. If we now choose $\alpha = 50$ and $\gamma = 10$, we have a new NITs scale on which

$$B = 50 + 10\,(b - 1.3)/1.9 \qquad [8.5.2]$$
$$= 50 - 10\,(1.3/1.9) + 10b/1.9$$
$$= 43.2 + 5.3b$$

and

$$D = 43.2 + 5.3d \quad .$$

Notice that with this scale definition, the normative mean of m = 1.3 logits becomes

$$M = 43.2 + 5.3\,(1.3) = 50 \text{ NITs}$$

If we now set b at

$$m + s = 1.3 + 1.9 = 3.2$$

then

$$M + S = 43.2 + 5.3\,(3.2) = 60 \text{ NITs}$$

so that

$$M + S - M = S = 60 - 50 = 10 \text{ NITs}$$

is the normative standard deviation on the new NIT scale.

Figure 8.5.1 shows the distribution of the 68 norming persons. Below their distribution are the ability measures for each score in logits and in NITs, and at the bottom are the KCTB items which define the variable. In Figure 8.5.1 we see that

30 NITs → m − 2s, 40 NITs → m − s, 50 NITs → m, 60 NITs → m + s and 70 NITs → m + 2s.

8.6. SUBSTANTIVE SCALING UNITS: SITS

We might instead choose to reference our new scale to substantive considerations such as a basal and a competency level, an entry and an exit level or some other two-position mastery hierarchy. To accomplish this we find the difficulty levels d_1 and d_2 on our logit scale which mark our choice of two criteria positions. Then we transform these logits to the values D_1 and D_2 on a new substantive scale or SIT which positions our criteria at easy to remember locations such as 50, 100 or 200.

If d_1 and d_2 identify the criteria positions in logits and D_1 and D_2 represent the desired easy to remember positions of these criteria on the new scale, then

$$\alpha = (D_1 d_2 - D_2 d_1)/(d_2 - d_1)$$
$$\gamma = (D_2 - D_1)/(d_2 - d_1)$$

and so

$$B = \alpha + \gamma b$$
$$D = \alpha + \gamma d$$

become

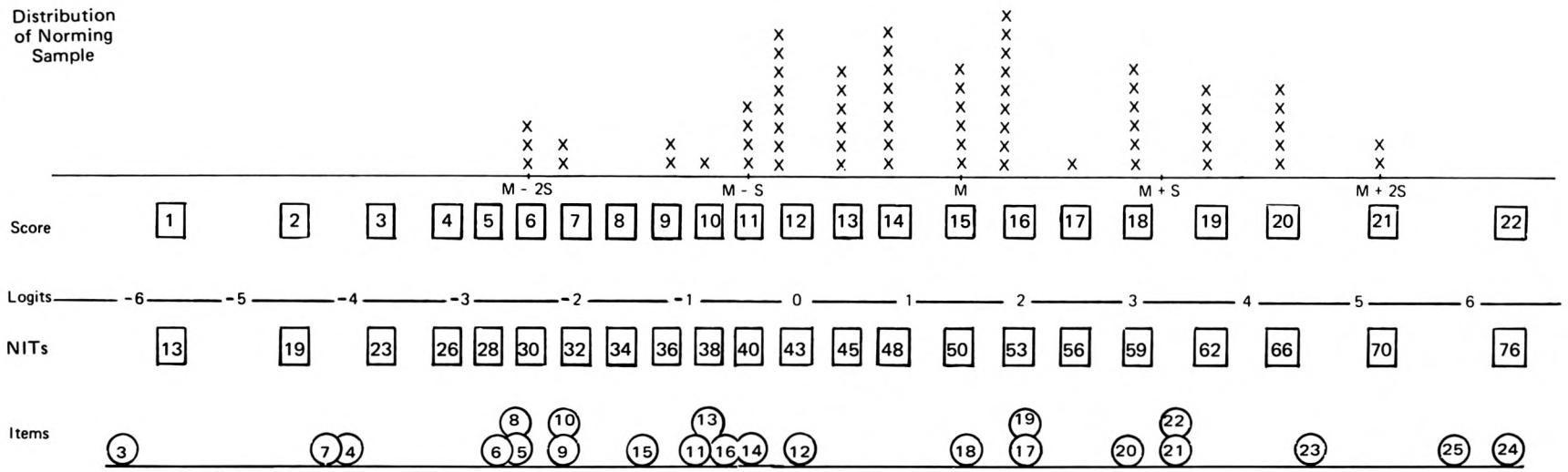
$$B = [(D_1 d_2 - D_2 d_1) + (D_2 - D_1)b]\,/(d_2 - d_1) \qquad [8.6.1]$$

and

$$D = [(D_1 d_2 - D_2 d_1) + (D_2 - D_1)d]\,/(d_2 - d_1) \quad .$$

FIGURE 8.5.1

SCALING THE KCTB IN NORMATIVE UNITS: NITS



In order to apply this method to the KCTB example, we will designate a basal level at the 3-tap median of $d_1 = -3.4$ logits and a competency level at the 5-tap median of $d_2 = 1.4$ logits. Then we will arrange to report these criteria at $D_1 = 30$ for basal and $D_2 = 50$ for competency using

$$\begin{aligned}\alpha &= [30(1.4) - 50(-3.4)]/[1.4 - (-3.4)] \\ &= (42 + 170)/4.8 \\ &= 44.2\end{aligned}$$

and

$$\begin{aligned}\gamma &= (50 - 30)/[(1.4 - (-3.4)] \\ &= 20/4.8 \\ &= 4.2\end{aligned}$$

which defines our new substantive scale of SITs as

$$B = 44.2 + 4.2b \qquad [8.6.2]$$

and

$$D = 44.2 + 4.2d$$

This scaling transforms the 3-tap median at $d_1 = -3.4$ logits to $D_1 = 44.2 + 4.2(-3.4) = 30$ SITs and the 5-tap median at $d_2 = 1.4$ logits to $D_2 = 44.2 + 4.2(1.4) = 50$ SITs.

In Figure 8.6.1 we show the KCTB items and a substantive definition of the KCT variable by marking the scale positions of each median number of taps. The ability scores in logits and in SITs are given below this substantive definition.

8.7 RESPONSE PROBABILITY SCALING UNITS: CHIPS

If we are interested in using our test to predict successful performance response rates, then a useful scale for these response probability units or CHIPs might be one that identified movement through the response probabilities of .10, .25, .50, .75 and .90 with easy to remember multiples along the variable like 5, 10, 20 or 25.

From the response model

$$p = \exp(b - d)/[1 + \exp(b - d)]$$

we can determine the differences $(b - d)$ between person ability and item difficulty which lead to the response probabilities .10, .25, .50, .75 and..90. Solving for $(b - d)$ in logits we have

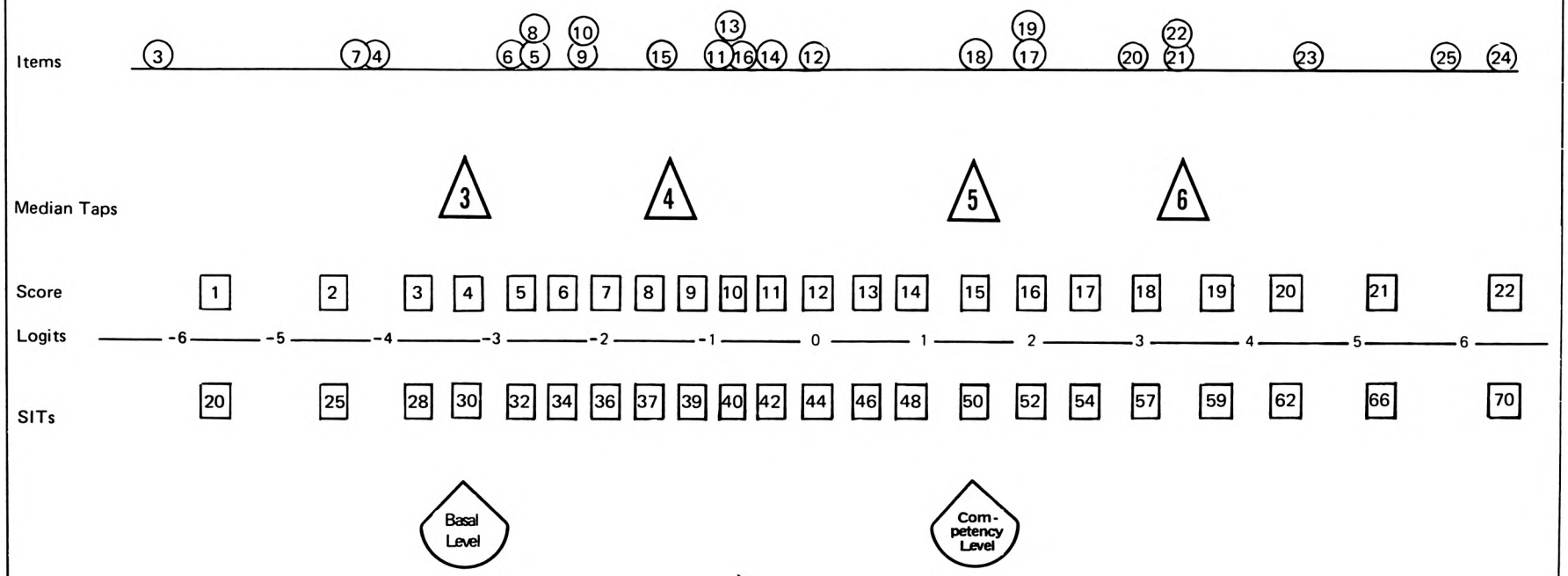
$$(b - d) = \ln [p/(1 - p)]$$

and hence

Probability of Success p	Difference Between Person Ability and Item Difficulty in Logits $b - d$
.10	-2.2
.25	-1.1
.50	0.0
.75	1.1
.90	2.2

FIGURE 8.6.1

SCALING THE KCTB IN SUBSTANTIVE UNITS: SITS



To determine a new scale in this manner we use

$$B = \alpha + \gamma (b - c) \qquad [8.7.1]$$

in which

c = either a normative or a substantive choice of location on the logit scale

γ = an appealing multiple of $1/1.1 = 0.91$ such as 5, 10, 20 or 25 leading to the γ values of 4.55, 9.1, 18.2 or 22.75.

and

α = 50, 100 or 500.

For KCTB we could make a normative choice of $c = 1.3$ logits at the logit mean of the norming group of 68 persons. We could also set $\gamma = 4.55$ giving us a CHIP spacing of 5 and choose α to locate the normative mean at 50. Then our CHIP scale formulation becomes

$$\begin{aligned} B &= 50 + 4.55 (b - 1.3) \qquad [8.7.2] \\ &= 44.09 + 4.55b \\ &= 44.1 + 4.6b \end{aligned}$$

Notice that when b is located at the mean of the norming group, then

$$B = 44.1 + 4.6 (1.3) = 50 \text{ CHIPs.}$$

Our choice of $\gamma = 4.55$ produces the following relations between the relative positions of a person at B and an item at D

Probability of Success p	Difference Between Person Ability and Item Difficulty in CHIPs B − D
.10	−10
.25	− 5
.50	0
.75	5
.90	10

Thus we expect that when any person confronts any item 10 CHIPs below their ability the probability for a successful response is about .90. At 5 CHIPs below, the predicted success rate is .75. On the other side, if an item is 5 CHIPs more difficult than the person is able, we expect the success rate to drop to .25 and, when the person is at a disadvantage of 10 CHIPs, we expect success only .10 of the time.

Were we to decide on a substantive choice of scale location, we could use the KCT 5-tap median of 1.4 logits as our reference location instead of the norming sample mean at 1.3 logits. Then our CHIP scale formulation would become

$$\begin{aligned} D &= 50 + 4.55(d - 1.4) \qquad [8.7.3] \\ &= 43.63 + 4.55d \\ &= 43.6 + 4.6d \end{aligned}$$

and so

$$B = 43.6 + 4.6b$$

Now when b is at the 5-tap median of 1.4 logits then

$$B = 43.6 + 4.6(1.4) = 50 \text{ CHIPs.}$$

Table 8.7.1 brings together the logit, NIT, SIT and CHIP scales for the KCTB test.

TABLE 8.7.1

KCTB SCORES, MEASURES AND ERRORS IN LOGITS, NITS, SITS AND CHIPS

Test Score	Person Ability				Measurement Error			
	Logits	NITs	SITs	CHIPs	Logits	NITs	SITs	CHIPs
22	6.26	76	70	73	1.20	6	5	6
21	5.15	70	66	68	0.99	5	4	5
20	4.31	66	62	64	0.89	5	4	4
19	3.60	62	59	61	0.83	4	3	4
18	2.99	59	57	58	0.78	4	3	4
17	2.42	56	54	55	0.76	4	3	3
16	1.88	53	52	53	0.75	4	3	3
15	1.35	50	50	50	0.74	4	3	3
14	0.84	48	48	48	0.73	4	3	3
13	0.35	45	46	46	0.70	4	3	3
12	-0.10	43	44	44	0.67	4	3	3
11	-0.51	40	42	42	0.65	3	3	3
10	-0.90	38	40	40	0.63	3	3	3
9	-1.28	36	39	38	0.62	3	3	3
8	-1.65	34	37	37	0.62	3	3	3
7	-2.02	32	36	35	0.63	3	3	3
6	-2.42	30	34	33	0.65	3	3	3
5	-2.85	28	32	31	0.69	4	3	3
4	-3.33	26	30	29	0.74	4	3	3
3	-3.90	23	28	26	0.82	4	3	4
2	-4.65	19	25	23	0.95	5	4	4
1	-5.75	13	20	18	1.23	7	5	6

[8.5.2] NITs $B = 43.2 + 5.3b$, $SE(B) = 5.3\ SE(b)$

[8.6.2] SITs $B = 44.2 + 4.2b$, $SE(B) = 4.2\ SE(b)$

[8.7.2] CHIPs $B = 44.1 + 4.6b$, $SE(B) = 4.6\ SE(b)$

8.8 REPORTING FORMS

The use of the Rasch model in test construction can facilitate test interpretation. We illustrate this with a reporting form developed for the KCT variable.

Figure 8.8.1 provides a map of the KCT variable. This map shows all of the data gathered thus far: the KCTB items positioned along the variable by their difficulty levels, the substantive criteria of number of taps, reverses and distance across blocks and the normative information of median ages for children and mean and standard deviation for adults. The map shows the extent to which the KCT variable has been defined and how various possible KCT measures relate to substantive and normative considerations. A KCT report form can be developed from this map.

Figure 8.8.2 is a report form for interpreting individual performance on the KCTB. This form, which could be used for a single individual or an entire class, shows the performance of Persons 12M and 88M as well as a response record identical in score to 88M but designed to show a "sleeping" pattern of several unexpected failures.

Notice that Person 12M with his score of 5 is located at -2.8 logits on the KCT variable. This puts him halfway between 3 and 4 taps substantively and at the 5 year old median normatively. Person 88M, however, at 3.0 logits is functioning at 6 taps substantively and at about one standard deviation above the adult mean normatively.

In many instances it will be useful to detect misfit immediately upon recording a person's responses. The report form in Figure 8.8.2 is ideal for this purpose. Once we have recorded the correct or incorrect response to each item at its position on the variable and also the consequent position of the person on the same variable. Misfit can be estimated directly from this completed answer form by means of a Misfit Ruler.

Figure 8.8.3 shows a Misfit Ruler scaled in logits. It is marked to indicate the logit deviations and a corresponding misfit index $y^2 = (z^2 - 1)$ to the left and right of its center. Notice that the unexpected response deviations of 1, 2, 3 and 4 logits indicate y^2's of 2, 6, 19 and 54 respectively. By positioning the center of the ruler at the point on the variable where the person is located and comparing the ruler's markings with the person's response to each item we can calculate, at a glance, the misfit of the person's record.

Whenever an unexpected response is observed, namely a "0" for an incorrect response in the easy region to the left or a "1" for a correct response in the hard region to the right, then the corresponding y^2's on the ruler are added to form their sum $Q = \Sigma\, y^2$ for just the unexpected pieces of the record. This sum Q divided by the square root of the total number of items L on the test yields the misfit statistic

$$U = Q/L^{1/2} \qquad [7.8.1]$$

which, if the record fits the response model, is distributed approximately.

$$U \sim N(0,2) \quad .$$

This easy to calculate statistic can be used to evaluate misfit. When $U > 5$ the probability that the record is acceptable has dropped below .01, and it seems reasonable to question the validity of the record.

In a practical application with a batch of records to evaluate, it is most reasonable to begin with the record for which U is maximum and to see if the source of invalidity

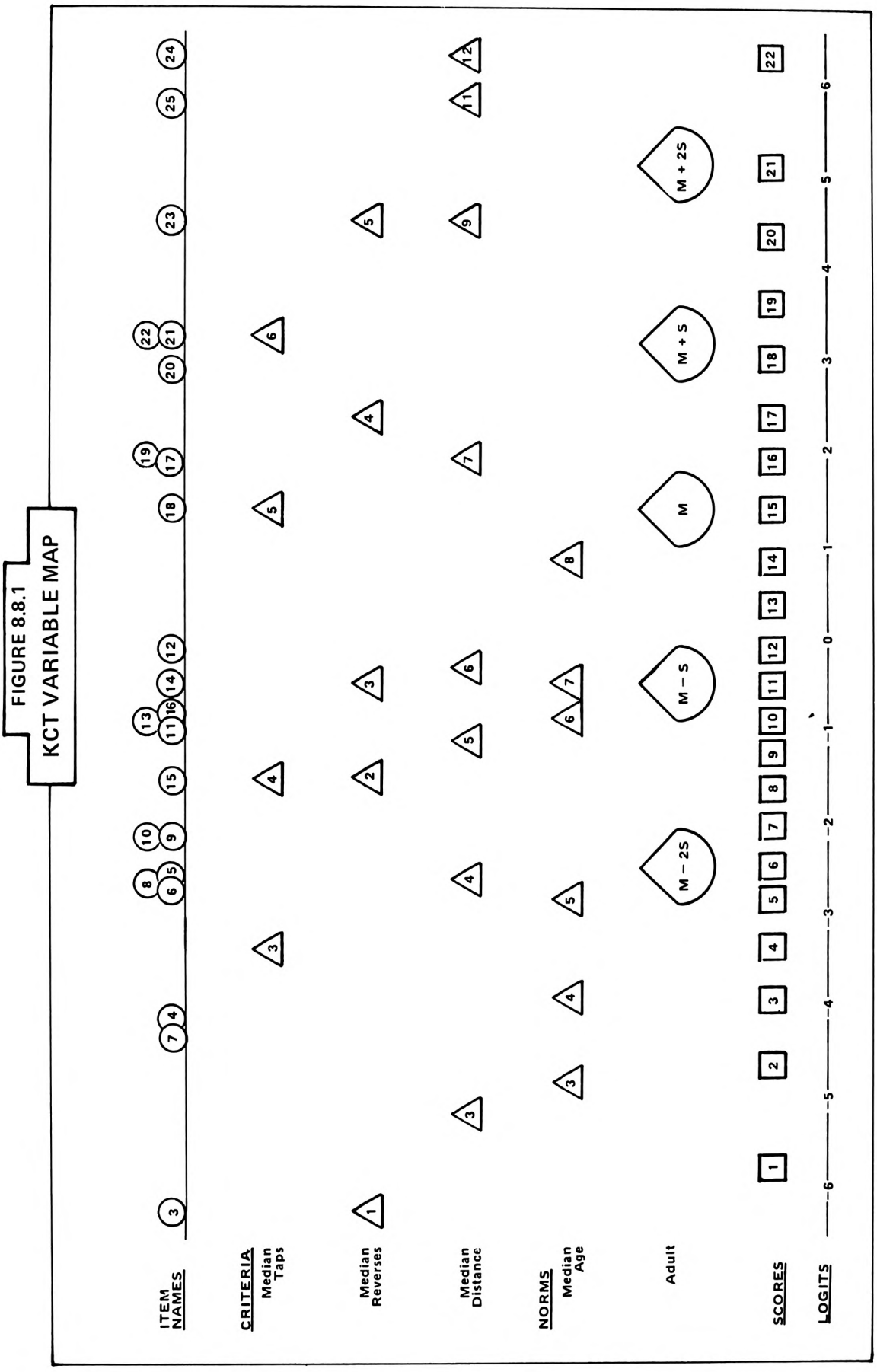
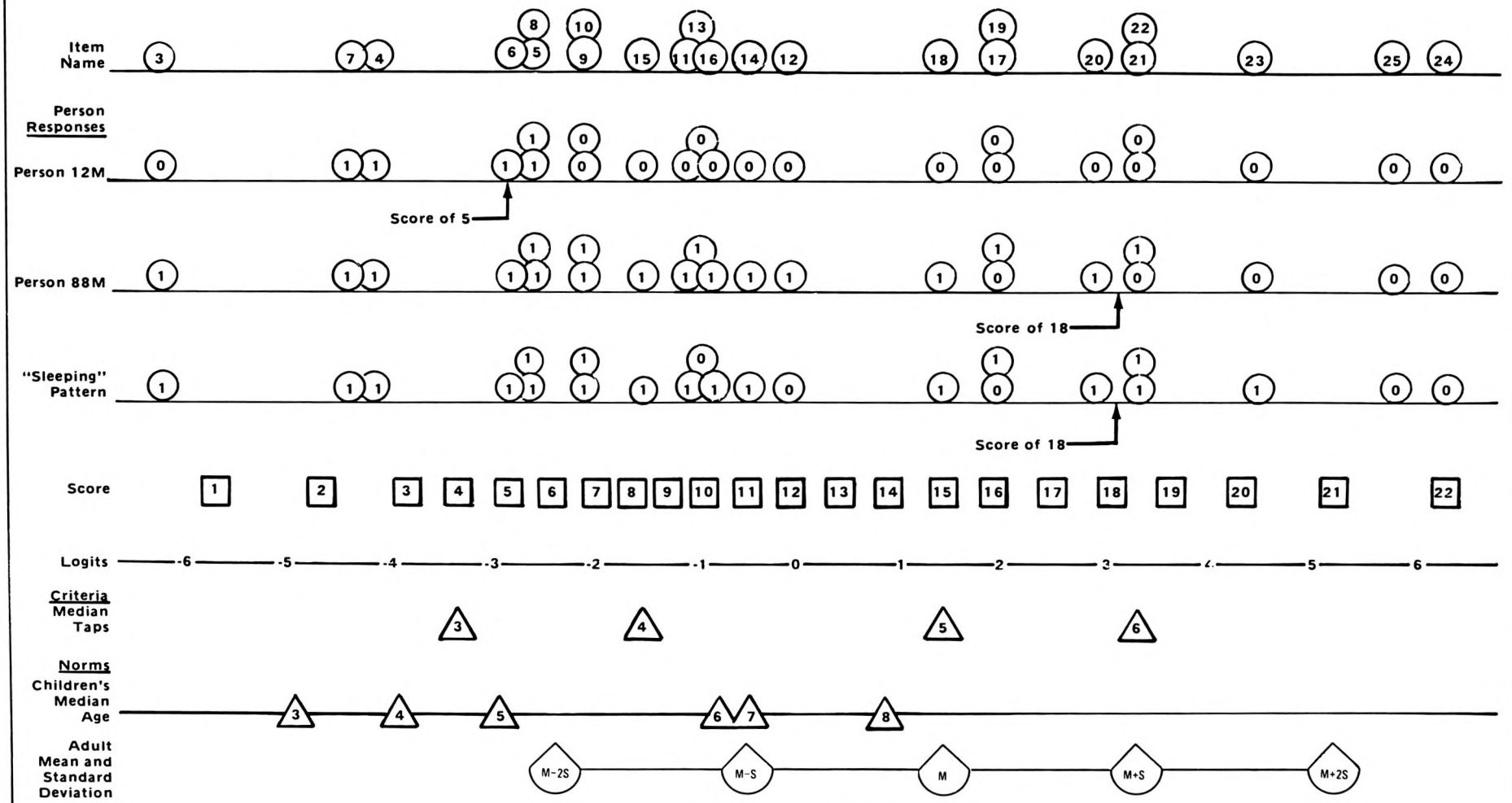

FIGURE 8.8.1
KCT VARIABLE MAP
ITEM NAMES
CRITERIA
Median Taps
Median Reverses
Median Distance
NORMS
Median Age
Adult
M − 2S
M − S
M
M + S
M + 2S
SCORES
LOGITS

FIGURE 8.8.2

KCTB TEST INTERPRETATION FORM



Items 1, 2, 26, 27 and 28 not included on this form

can be identified and dealt with. Subsequent cases can then be handled in order of U until all useful explanations of the invalidities implied by $U > 5$ are discovered.

Figure 8.8.2 shows the test segment of 23 KCTB items in order of increasing difficulty together with three response records. The response records show the correct and incorrect responses on the 23 items. To evaluate the fit of Person 12M's record the center of the Misfit Ruler is placed at the arrow marking his position at -2.8 logits determined by his score of 5. His incorrect response to Item 3 at -6.2 logits produces a y^2 of about 30 for Q = 30 and $U = 30/23^{1/2} = 6.3$. This corresponds to a t = 4.5 which is very close to the more exact value of t given in Table 7.8.3. Again, we see that this response is too improbable to be accepted as part of a valid measure of Person 12M.

The Misfit Ruler has also been applied to Person 88M and to the "Sleeping" pattern with the same score of 18. The pattern for Person 88M produces a response record that yields a $\Sigma y^2 = 2$, and U = 0.4. The "sleeping" patterns, however, produces a

$$\Sigma y^2 = 46 + 20 + 2 + 3 = 71$$

and so a

$$U = 71/23^{1/2} = 14.8 \quad .$$

These results are summarized in Table 8.8.1

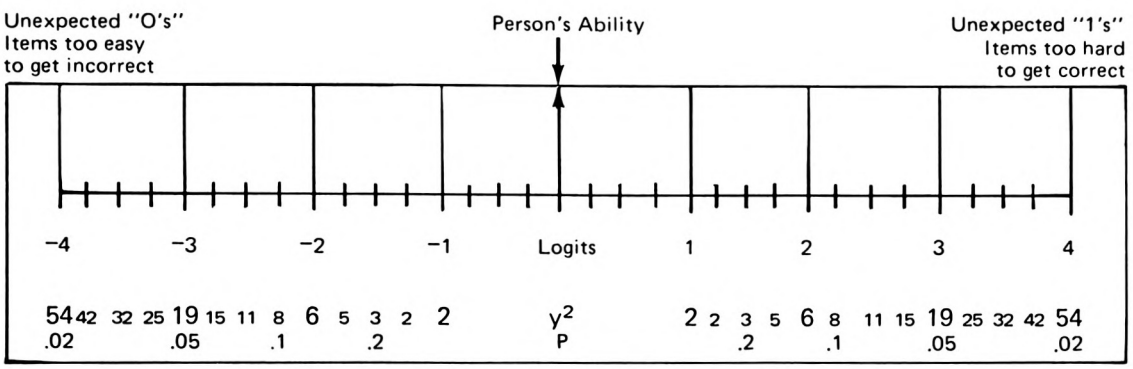
TABLE 8.8.1

QUICK ANALYSIS OF RESPONSE RECORD VALIDITY

Person	Score r	Sum of Unexpected Responses $Q = \Sigma y^2$	Fit Statistic $U = Q/L^{1/2}$
12M	5	30	6.3*
88M	18	2	0.4
"Sleeping" Pattern	18	46 + 20 + 2 + 3 = 71	14.8*
	L = 23		*Misfit

FIGURE 8.8.3

MISFIT RULER



HOW TO USE THE MISFIT RULER:

1. Position the items on metric record form corresponding to ruler metric.
2. Record person's responses to items on record form.
3. Locate person's ability position on record form by counting score r and positioning it between the rth and the (r + 1)th item locations.
4. Place center of Misfit Ruler at person's ability position.
5. Sum y^2 for all unexpected responses, "0's" to the left and "1's" to the right to form Q.
6. Let L equal the total number of items.
7. Calculate misfit statistic $U = Q / L^{1/2}$.
8. If $U > 5$ examine the person's record further for sources of invalidity.

P is the improbability of each response.

APPENDICES

TABLE A	211 & 212
TABLE B	213 & 214
TABLE C	215

TABLE A

RELATIVE ABILITY x_{fw} FOR UNIFORM TESTS IN LOGITS

Relative Score f > .50	Test Width w 1	2	3	4	5	6	7	8	Relative Score f < .50
.50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.50
.51	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	.49
.52	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	.48
.53	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	.47
.54	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	.46
.55	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	.45
.56	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.5	.44
.57	0.3	0.3	0.3	0.4	0.4	0.5	0.5	0.6	.43
.58	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.7	.42
.59	0.4	0.4	0.4	0.5	0.5	0.6	0.7	0.7	.41
.60	0.4	0.4	0.5	0.5	0.6	0.7	0.7	0.8	.40
.61	0.5	0.5	0.5	0.6	0.7	0.7	0.8	0.9	.39
.62	0.5	0.5	0.6	0.6	0.7	0.8	0.9	1.0	.38
.63	0.5	0.6	0.6	0.7	0.8	0.9	1.0	1.1	.37
.64	0.6	0.6	0.7	0.7	0.8	0.9	1.1	1.2	.36
.65	0.6	0.7	0.7	0.8	0.9	1.0	1.1	1.3	.35
.66	0.7	0.7	0.8	0.9	1.0	1.1	1.2	1.3	.34
.67	0.7	0.8	0.8	0.9	1.0	1.2	1.3	1.4	.33
.68	0.8	0.8	0.9	1.0	1.1	1.2	1.4	1.5	.32
.69	0.8	0.9	0.9	1.0	1.2	1.3	1.4	1.6	.31
.70	0.9	0.9	1.0	1.1	1.2	1.4	1.5	1.7	.30
.71	0.9	1.0	1.0	1.2	1.3	1.4	1.6	1.8	.29
.72	1.0	1.0	1.1	1.2	1.4	1.5	1.7	1.9	.28
.73	1.0	1.1	1.2	1.3	1.4	1.6	1.8	2.0	.27
.74	1.1	1.1	1.2	1.3	1.5	1.7	1.9	2.1	.26
.75	1.1	1.2	1.3	1.4	1.6	1.7	1.9	2.1	.25
.76	1.2	1.2	1.3	1.5	1.6	1.8	2.0	2.2	.24
.77	1.2	1.3	1.4	1.5	1.7	1.9	2.1	2.3	.23
.78	1.3	1.4	1.5	1.6	1.8	2.0	2.2	2.4	.22
.79	1.3	1.4	1.5	1.7	1.9	2.1	2.3	2.5	.21
.80	1.4	1.5	1.6	1.8	1.9	2.2	2.4	2.6	.20
.81	1.5	1.6	1.7	1.8	2.0	2.2	2.5	2.7	.19
.82	1.5	1.6	1.7	1.9	2.1	2.3	2.6	2.8	.18
.83	1.6	1.7	1.8	2.0	2.2	2.4	2.7	2.9	.17
.84	1.7	1.8	1.9	2.1	2.3	2.5	2.8	3.0	.16
.85	1.8	1.8	2.0	2.2	2.4	2.6	2.9	3.2	.15
.86	1.8	1.9	2.1	2.3	2.5	2.7	3.0	3.3	.14
.87	1.9	2.0	2.2	2.4	2.6	2.8	3.1	3.4	.13
.88	2.0	2.1	2.3	2.5	2.7	2.9	3.2	3.5	.12
.89	2.1	2.2	2.4	2.6	2.8	3.1	3.3	3.7	.11
.90	2.2	2.3	2.5	2.7	2.9	3.2	3.5	3.8	.10
.91	2.3	2.4	2.6	2.8	3.1	3.3	3.6	3.9	.09
.92	2.5	2.6	2.7	2.9	3.2	3.5	3.8	4.1	.08
.93	2.6	2.7	2.9	3.1	3.4	3.6	4.0	4.3	.07
.94	2.8	2.9	3.1	3.3	3.5	3.8	4.1	4.5	.06
.95	3.0	3.1	3.3	3.5	3.7	4.0	4.4	4.7	.05
.96	3.2	3.3	3.5	3.7	4.0	4.3	4.6	5.0	.04
.97	3.5	3.6	3.8	4.0	4.3	4.6	5.0	5.3	.03
.98	3.9	4.0	4.2	4.5	4.7	5.1	5.4	5.8	.02
.99	4.6	4.8	4.9	5.2	5.5	5.8	6.1	6.5	.01
f > .50 Measure $b_f = h + x_{fw}$									f < .50 Measure $b_f = h - x_{fw}$

Test Score: r Test Length: L Relative Score: f = r/L

Test Height: $h = \sum_{i}^{L} d_i/L$ Test Width: $w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$

TABLE A
RELATIVE ABILITY x_{fw} FOR UNIFORM TESTS IN LOGITS (Continued)

Relative Score f > .50	Test Width w: 8	9	10	11	12	13	14	15	Relative Score f < .50
.50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.50
.51	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	.49
.52	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	.48
.53	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.5	.47
.54	0.3	0.4	0.4	0.4	0.5	0.5	0.6	0.6	.46
.55	0.4	0.5	0.5	0.6	0.6	0.7	0.7	0.8	.45
.56	0.5	0.6	0.6	0.7	0.7	0.8	0.8	0.9	.44
.57	0.6	0.6	0.7	0.8	0.8	0.9	1.0	1.1	.43
.58	0.7	0.7	0.8	0.9	1.0	1.0	1.1	1.2	.42
.59	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	.41
.60	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	.40
.61	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.7	.39
.62	1.0	1.1	1.2	1.3	1.4	1.6	1.7	1.8	.38
.63	1.1	1.2	1.3	1.4	1.6	1.7	1.8	2.0	.37
.64	1.2	1.3	1.4	1.6	1.7	1.8	2.0	2.1	.36
.65	1.3	1.4	1.5	1.7	1.8	2.0	2.1	2.3	.35
.66	1.3	1.5	1.6	1.8	1.9	2.1	2.2	2.4	.34
.67	1.4	1.6	1.7	1.9	2.1	2.2	2.4	2.6	.33
.68	1.5	1.7	1.8	2.0	2.2	2.4	2.5	2.7	.32
.69	1.6	1.8	1.9	2.1	2.3	2.5	2.7	2.9	.31
.70	1.7	1.9	2.1	2.2	2.4	2.6	2.8	3.0	.30
.71	1.8	2.0	2.2	2.4	2.6	2.8	3.0	3.2	.29
.72	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	.28
.73	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.5	.27
.74	2.1	2.3	2.5	2.7	2.9	3.2	3.4	3.6	.26
.75	2.1	2.4	2.6	2.8	3.1	3.3	3.5	3.8	.25
.76	2.2	2.5	2.7	2.9	3.2	3.4	3.7	3.9	.24
.77	2.3	2.6	2.8	3.1	3.3	3.6	3.8	4.1	.23
.78	2.4	2.7	2.9	3.2	3.4	3.7	4.0	4.2	.22
.79	2.5	2.8	3.0	3.3	3.6	3.8	4.1	4.4	.21
.80	2.6	2.9	3.1	3.4	3.7	4.0	4.3	4.6	.20
.81	2.7	3.0	3.3	3.5	3.8	4.1	4.4	4.7	.19
.82	2.8	3.1	3.4	3.7	4.0	4.3	4.6	4.9	.18
.83	2.9	3.2	3.5	3.8	4.1	4.4	4.7	5.0	.17
.84	3.0	3.3	3.6	3.9	4.2	4.6	4.9	5.2	.16
.85	3.2	3.4	3.8	4.1	4.4	4.7	5.0	5.4	.15
.86	3.3	3.6	3.9	4.2	4.5	4.9	5.2	5.5	.14
.87	3.4	3.7	4.0	4.3	4.7	5.0	5.4	5.7	.13
.88	3.5	3.8	4.2	4.5	4.8	5.2	5.5	5.9	.12
.89	3.7	4.0	4.3	4.6	5.0	5.3	5.7	6.1	.11
.90	3.8	4.1	4.5	4.8	5.2	5.5	5.9	6.3	.10
.91	3.9	4.3	4.6	5.0	5.3	5.7	6.1	6.5	.09
.92	4.1	4.4	4.8	5.2	5.5	5.9	6.3	6.7	.08
.93	4.3	4.6	5.0	5.4	5.7	6.1	6.5	6.9	.07
.94	4.5	4.8	5.2	5.6	5.9	6.3	6.7	7.1	.06
.95	4.7	5.1	5.4	5.8	6.2	6.6	7.0	7.4	.05
.96	5.0	5.3	5.7	6.1	6.5	6.9	7.3	7.7	.04
.97	5.3	5.7	6.1	6.4	6.8	7.2	7.7	8.1	.03
.98	5.8	6.1	6.5	6.9	7.3	7.7	8.1	8.6	.02
.99	6.5	6.9	7.3	7.7	8.1	8.5	8.9	9.3	.01
f > .50 Measure $b_f = h + x_{fw}$									f < .50 Measure $b_f = h - x_{fw}$

Test Score: r Test Length: L Relative Score: f = r/L

Test Height: $h = \sum_{i}^{L} d_i/L$ Test Width: $w = [(d_L + d_{L-1} - d_2 - d_1)/2][L/(L-2)]$

TABLE B

ERROR COEFFICIENT $C_{fw}^{1/2}$ FOR UNIFORM TESTS IN LOGITS

Relative Score f > .50	Test Width w 1	2	3	4	5	6	7	8	Relative Score f < .50
.50	2.0	2,1	2.2	2.3	2.4	2.6	2.7	2.9	.50
.51	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.49
.52	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.48
.53	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.47
.54	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.46
.55	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.45
.56	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.44
.57	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.43
.58	2.0	2.1	2.2	2.3	2.4	2.6	2.7	2.9	.42
.59	2.1	2.1	2.2	2.3	2.5	2.6	2.7	2.9	.41
.60	2.1	2.1	2.2	2.3	2.5	2.6	2.7	2.9	.40
.61	2.1	2.1	2.2	2.3	2.5	2.6	2.8	2.9	.39
.62	2.1	2.1	2.2	2.3	2.5	2.6	2.8	2.9	.38
.63	2.1	2.1	2.2	2.4	2.5	2.6	2.8	2.9	.37
.64	2.1	2.2	2.2	2.4	2.5	2.6	2.8	2.9	.36
.65	2.1	2.2	2.3	2.4	2.5	2.6	2.8	2.9	.35
.66	2.1	2.2	2.3	2.4	2.5	2.6	2.8	2.9	.34
.67	2.1	2.2	2.3	2.4	2.5	2.7	2.8	2.9	.33
.68	2.2	2.2	2.3	2.4	2.5	2.7	2.8	3.0	.32
.69	2.2	2.2	2.3	2.4	2.6	2.7	2.8	3.0	.31
.70	2.2	2.3	2.3	2.4	2.6	2.7	2.8	3.0	.30
.71	2.2	2.3	2.4	2.5	2.6	2.7	2.8	3.0	.29
.72	2.2	2.3	2.4	2.5	2.6	2.7	2.9	3.0	.28
.73	2.3	2.3	2.4	2.5	2.6	2.7	2.9	3.0	.27
.74	2.3	2.3	2.4	2.5	2.6	2.8	2.9	3.0	.26
.75	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	.25
.76	2.4	2.4	2.5	2.6	2.7	2.8	2.9	3.1	.24
.77	2.4	2.4	2.5	2.6	2.7	2.8	3.0	3.1	.23
.78	2.4	2.5	2.6	2.6	2.8	2.9	3.0	3.1	.22
.79	2.5	2.5	2.6	2.7	2.8	2.9	3.0	3.1	.21
.80	2.5	2.6	2.6	2.7	2.8	2.9	3.1	3.2	.20
.81	2.6	2.6	2.7	2.8	2.9	3.0	3.1	3.2	.19
.82	2.6	2.7	2.7	2.8	2.9	3.0	3.1	3.2	.18
.83	2.7	2.7	2.8	2.9	3.0	3.1	3.2	3.3	.17
.84	2.7	2.8	2.9	2.9	3.0	3.1	3.2	3.3	.16
.85	2.8	2.9	2.9	3.0	3.1	3.2	3.3	3.4	.15
.86	2.9	2.9	3.0	3.1	3.2	3.3	3.4	3.4	.14
.87	3.0	3.0	3.1	3.2	3.2	3.3	3.4	3.5	.13
.88	3.1	3.1	3.2	3.3	3.3	3.4	3.5	3.6	.12
.89	3.2	3.2	3.3	3.4	3.4	3.5	3.6	3.7	.11
.90	3.3	3.4	3.4	3.5	3.6	3.7	3.7	3.8	.10
.91	3.5	3.5	3.6	3.7	3.7	3.8	3.9	3.9	.09
.92	3.7	3.7	3.8	3.8	3.9	4.0	4.0	4.1	.08
.93	3.9	4.0	4.0	4.1	4.1	4.2	4.3	4.3	.07
.94	4.2	4.2	4.3	4.3	4.4	4.5	4.5	4.6	.06
.95	4.6	4.6	4.7	4.7	4.8	4.8	4.9	4.9	.05
.96	5.1	5.1	5.2	5.2	5.3	5.3	5.4	5.4	.04
.97	5.9	5.9	5.9	6.0	6.0	6.0	6.1	6.1	.03
.98	7.1	7.2	7.2	7.2	7.3	7.3	7.3	7.4	.02
.99	10.1	10.1	10.1	10.1	10.1	10.2	10.2	10.2	.01
f > .50									f < .50

Test Score: r Test Length: L Relative Score: f = r/L

Test Height: $h = \sum_{i}^{L} d_i/L$ Test Width: $w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$

Standard Error: $s_{fw} = C_{fw}^{1/2}/L^{1/2}$

TABLE B
ERROR COEFFICIENT $C_{fw}^{1/2}$ FOR UNIFORM TESTS IN LOGITS (Continued)

Relative Score $f > .50$	Test Width w								Relative Score $f < .50$
	8	9	10	11	12	13	14	15	
.50	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.50
.51	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.49
.52	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.48
.53	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.47
.54	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.46
.55	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.45
.56	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.44
.57	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.43
.58	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.42
.59	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.41
.60	2.9	3.0	3.2	3.3	3.5	3.6	3.7	3.9	.40
.61	2.9	3.1	3.2	3.3	3.5	3.6	3.8	3.9	.39
.62	2.9	3.1	3.2	3.3	3.5	3.6	3.8	3.9	.38
.63	2.9	3.1	3.2	3.3	3.5	3.6	3.8	3.9	.37
.64	2.9	3.1	3.2	3.4	3.5	3.6	3.8	3.9	.36
.65	2.9	3.1	3.2	3.4	3.5	3.6	3.8	3.9	.35
.66	2.9	3.1	3.2	3.4	3.5	3.6	3.8	3.9	.34
.67	2.9	3.1	3.2	3.4	3.5	3.6	3.8	3.9	.33
.68	3.0	3.1	3.2	3.4	3.5	3.6	3.8	3.9	.32
.69	3.0	3.1	3.2	3.4	3.5	3.6	3.8	3.9	.31
.70	3.0	3.1	3.2	3.4	3.5	3.6	3.8	3.9	.30
.71	3.0	3.1	3.3	3.4	3.5	3.6	3.8	3.9	.29
.72	3.0	3.1	3.3	3.4	3.5	3.7	3.8	3.9	.28
.73	3.0	3.1	3.3	3.4	3.5	3.7	3.8	3.9	.27
.74	3.0	3.2	3.3	3.4	3.5	3.7	3.8	3.9	.26
.75	3.0	3.2	3.3	3.4	3.6	3.7	3.8	3.9	.25
.76	3.1	3.2	3.3	3.4	3.6	3.7	3.8	3.9	.24
.77	3.1	3.2	3.3	3.5	3.6	3.7	3.8	3.9	.23
.78	3.1	3.2	3.4	3.5	3.6	3.7	3.8	3.9	.22
.79	3.1	3.3	3.4	3.5	3.6	3.7	3.8	4.0	.21
.80	3.2	3.3	3.4	3.5	3.6	3.7	3.9	4.0	.20
.81	3.2	3.3	3.4	3.5	3.7	3.8	3.9	4.0	.19
.82	3.2	3.4	3.5	3.6	3.7	3.8	3.9	4.0	.18
.83	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	.17
.84	3.3	3.4	3.5	3.6	3.7	3.9	4.0	4.1	.16
.85	3.4	3.5	3.6	3.7	3.8	3.9	4.0	4.1	.15
.86	3.4	3.5	3.6	3.7	3.8	3.9	4.0	4.1	.14
.87	3.5	3.6	3.7	3.8	3.9	4.0	4.1	4.2	.13
.88	3.6	3.7	3.8	3.9	4.0	4.1	4.1	4.2	.12
.89	3.7	3.8	3.9	4.0	4.0	4.1	4.2	4.3	.11
.90	3.8	3.9	4.0	4.1	4.1	4.2	4.3	4.4	.10
.91	3.9	4.0	4.1	4.2	4.3	4.3	4.4	4.5	.09
.92	4.1	4.2	4.3	4.3	4.4	4.5	4.6	4.6	.08
.93	4.3	4.4	4.5	4.5	4.6	4.7	4.7	4.8	.07
.94	4.6	4.6	4.7	4.8	4.8	4.9	5.0	5.0	.06
.95	4.9	5.0	5.0	5.1	5.2	5.2	5.3	5.3	.05
.96	5.4	5.5	5.5	5.6	5.6	5.7	5.7	5.8	.04
.97	6.1	6.2	6.2	6.3	6.3	6.3	6.4	6.4	.03
.98	7.4	7.4	7.4	7.5	7.5	7.5	7.6	7.6	.02
.99	10.2	10.2	10.3	10.3	10.3	10.3	10.4	10.4	.01
$f > .50$									$f < .50$

Test Score: r Test Length: L Relative Score: $f = r/L$

Test Height: $h = \sum_{i}^{L} d_i/L$ Test Width: $w = [(d_L + d_{L-1} - d_2 - d_1)/2]\,[L/(L-2)]$

Standard Error: $s_{fw} = C_{fw}^{1/2}/L^{1/2}$

TABLE C
MISFIT STATISTICS

Difference Between Person Ability and Item Difficulty (b−d)	Squared Standardized Residual $z^2=\exp(b-d)$	Improbability of the Response $p=1/(1+z^2)$	Relative Efficiency of the Observation $I=400p(1-p)$	Number of Items Needed To Maintain Equal Precision $L=1000/I$
−0.6, 0.3	1	.50	100	10
0.4, 0.8	2	.33	90	11
0.9, 1.2	3	.25	75	13
1.3, 1.4	4	.20	65	15
1.5, 1.4	5	.17	55	18
1.7, 1.8	6	.14	50	20
1.9, 2.0	7	.12	45	22
2.1	8	.11	40	25
2.2	9	.10	36	28
2.3	10	.09	33	30
2.4	11	.08	31	32
2.5	12	.08	28	36
2.6	14	.07	25	40
2.7	15	.06	23	43
2.8	17	.06	21	48
2.9	18	.05	20	50
3.0	20	.05	18	55
3.1	22	.04	16	61
3.2	25	.04	15	66
3.3	27	.04	14	73
3.4	30	.03	12	83
3.5	33	.03	11	91
3.6	37	.03	10	100
3.7	41	.02	9	106
3.8	45	.02	9	117
3.9	50	.02	8	129
4.0	55	.02	7	142
4.1	60	.02	6	156
4.2	67	.02	6	172
4.3	74	.01	5	189
4.4	81	.01	5	209
4.5	90	.01	4	230
4.6	99	.01	4	254

REFERENCES

Allerup, P. and Sorber, G. *The Rasch Model for Questionnaires With a Computer Program.* Copenhagen: Danmarks Paedagogiske Institut, 1977.

Andersen, E. B. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society B*, 1970, *32*, 283-301.

Andersen, E. B. Asymptotic properties of conditional likelihood ratio tests. *Journal of the American Statistical Association*, 1971, *66*, 630-633.

Andersen, E. B. The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society B*, 1972a, *34*, 42-54.

Andersen, E. B. A computer program for solving a set of conditional maximum likelihood equations arising in the Rasch model for questionnaires. *Research Memorandum 72-6*. Princeton, N.J.: Educational Testing Service, 1972b.

Andersen, E. B. *Conditional Inference and Models for Measuring.* Copenhagen: Mentalhygiejnisk Forlag, 1973.

Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika*, 1977, *42*, 69-81.

Andrich, D. Latent trait psychometric theory in the measurement and evaluation of essay writing ability. Doctoral dissertation, University of Chicago, 1973.

Andrich, D. The Rasch multiplicative binomial model: Applications to attitude data. *Research Report No. 1.* Measurement and Statistics Laboratory, Department of Education, University of Western Australia, 1975.

Angoff, W. H. Measurement and scaling. In C. W. Harris (Ed.), *Encyclopedia of Educational Research.* New York: Macmillan, 1960.

Arthur, Grace. *A Point Scale of Performance Tests.* New York: Psychological Corporation, 1947.

Baker, F. B. UNIVAC scientific computer program for test scoring and item analysis. *Behavioral Sciences*, 1959, *4*, 254-255.

Baker, F. B. Empirical comparison of item parameters based on the logistic and normal functions, *Psychometrika*, 1961, *26*, 235-246.

Baker, F. B. Generalized item and test analysis program—a program for the Control Data 1604 Computer. *Educational and Psychological Measurement*, 1963, *23*, 187-190.

Barndorff-Nielsen, O. *Information and Exponential Families in Statistical Theory.* New York: John Wiley and Sons, 1978.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. Novick, *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley, 1968.

Choppin, B. An item bank using sample-free calibration. *Nature*, 1968, *219*, 870-872.

Choppin, B. *Item Banking and the Monitoring of Achievement.* Slough, England: National Foundation for Educational Research, 1978.

Cohen, L. A modified logistic response model for item analysis. Manuscript, 1976.

Connolly, A. J., Nachtman, W. and Pritchett, E. M. *Keymath: Diagnostic Arithmetic Test.* Circle Pines, Minn.: American Guidance Service, 1971.

Cornish, G. and Wines, R. *ACER Mathematics Profile Series (MAPS).* Hawthorn, Victoria: Australian Council for Educational Research, 1977.

Douglas, Graham A. Test design strategies for the Rasch psychometric model. Doctoral dissertation, University of Chicago, 1974.

Draba, R. E. The Rasch model and legal criteria of a "reasonable" classification. Doctoral dissertation, University of Chicago, 1978.

Elliott, C. D., Murray, D. C., and Pearson, L. S. *The British Ability Scales.* Slough, England: National Foundation for Educational Research, 1977.

Fischer, G. H. and Scheiblechner, H. H. Two simple methods for asymptotically unbiased estimation in Rasch's model with two categories of answers. *Research Bulletin No. 1*, Psychological Institute, University of Vienna, 1970.

Gulliksen, H. *Theory of Mental Tests.* New York: John Wiley & Sons, 1950.

Gustafsson, J. E. The Rasch model for dichotomous items: Theory, applications and a computer program. *Report No. 63.* Institute of Education, University of Goteberg, 1977.

Habermann, S. Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 1977, *5*, 815-841.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 1947, *61*.

Loevinger, J. Person and population as psychometric concepts. *Psychological Review*, 1965, *72*, 143-155.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Mason, G. P. and Odeh, R. E. A short-cut formula for standard deviation. *Journal of Educational Measurement*, 1968, *5*, 319-320.

Mead, R. J. Analysis of fit to the Rasch model. Doctoral dissertation, University of Chicago, 1975.

Panchapakesan, N. The simple logistic model and mental measurement. Doctoral dissertation, University of Chicago, 1969.

Rasch, G. On simultaneous factor analysis in several populations. In *Uppsala Symposium on Psychological Factor Analysis.* Stockholm: Almquist and Wiksells, 1953, 65-71.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedagogiske Institut, 1960 (To be reprinted by University of Chicago Press, 1980).

Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1961, *4*, 321-333.

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, 1966a, 89-108.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology.* 1966b, *19*, 49-57.

Rasch, G. An informal report on the present state of a theory of objectivity in comparisons. In L. J. van der Kamp and C. A. J. Viek (Eds.), *Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof."* Leiden, 1967.

Rasch, G. A mathematical theory of objectivity and its consequences for model construction. In *Report from European Meeting on Statistics, Econometrics and Management Sciences*, Amsterdam, 1968.

Rentz, R. R. and Bashaw, W. L. *Equating Reading Tests with the Rasch Model.* Athens, Georgia: Educational Resource Laboratory, 1975.

Rentz, R. R. and Bashaw, W. L. The national reference scale for reading: An application of the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 161-180.

Thorndike, E. L., *et al. The Measurement of Intelligence.* New York: Columbia University, Teachers College, 1926.

Thurstone, L. L. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 1925, *16*, 433-451.

Thurstone, L. L. The unit of measurement in educational scales. *Journal of Educational Psychology*, 1927, *18*, 505-524.

Tucker, L. R. Scales minimizing the importance of reference groups. In *Proceedings of the 1952 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1953.

Willmott, A. and Fowles, D. *The Objective Interpretation of Test Performance: The Rasch Model Applied.* Atlantic Highlands, N.J.: NFER Publishing Co., Ltd., 1974.

Woodcock, R. W. *Woodcock Reading Mastery Tests.* Circle Pines, Minnesota: American Guidance Service, 1974.

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 97-116.

Wright, B. D. and Douglas, G. A. Best test design and self-tailored testing. *Research Memorandum No. 19*, Statistical Laboratory, Department of Education, University of Chicago, 1975a.

Wright, B. D. and Douglas, G. A. Better procedures for sample-free item analysis. *Research Memorandum No. 20*, Statistical Laboratory, Department of Education, University of Chicago, 1975b.

Wright, B. D. and Douglas, G. A. Rasch item analysis by hand. *Research Memorandum No. 21*, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D. and Douglas, G. A. Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1977a, *1*, 281-294.

Wright, B. D. and Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, 1977b, *37*, 573-586.

Wright, B. D. and Mead, R. J. CALFIT: Sample-free calibration with a Rasch measurement model. *Research Memorandum No. 18.* Statistical Laboratory, Department of Education, University of Chicago, 1975.

Wright, B. D. and Mead, R. J. BICAL: Calibrating items with the Rasch model. *Research Memorandum No. 23*, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D. and Mead, R. J. *The Use of Measurement Models in the Definition and Application of Social Science Variables.* Arlington, VA: U.S. Army Research Institute Technical Report DAHC19-76-G-0011, 1977.

Wright, B. D., Mead, R. J. and Draba, R. E. Detecting and correcting test item bias with a logistic response model. *Research Memorandum No. 22*, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis, *Educational and Psychological Measurement*, 1969, *29*, 23-48.

INDEX

Ability β and b (*see* Person ability measure)
Additive scale factor α (*see* Scale additive factor)
Analysis of fit (*see* Fit)

Bank (*see* Item bank building)
Best test design (*see* Test design)
Beta β (*see* Person ability measure)
BICAL, 46 - 54
  control, 46 - 47
  output, 46, 48 - 52, 54

Calibration δ and d (*see* Item difficulty calibration)
Chain, 99
Chicago probability unit (*see* Scale CHIP)
CHIP (*see* Scale CHIP)
Common item equating, 108 - 109, 112 - 118
Common person equating, 106 - 112
Computing algorithms:
  BICAL, 46 - 54
  PROX, 61 - 62
    hand example, 30 - 44
    computer example, 46 - 55
  UCON, 62 - 65
  UFORM, 143 - 151
    tables, 146, 212, 214
Correcting a measure, 181 - 190
Connecting two tests, 96 - 98 (*see* Linking test forms)
Control lines for identity plots, 94 - 95
Criterion referencing, 118 - 121, 199 - 202, 204, 206 - 207
Crude fit (*see* Fit)

Data matrix, 10, 18, 31, 33, 68, 107 - 109
Degrees of freedom, 23 - 24, 71, 74, 77, 79
  crude fit, 125
  item fit, 24, 77, 79
  link analysis, 96
  person fit, 23, 76 - 77, 79, 165 - 168
Delta δ (*see* Item difficulty calibration)
Design of best test (*see* Test design)
Diagnosing misfit, 170 - 180
Difficulty δ and d (*see* Item difficulty calibration)
Discrimination (*see* Item discrimination index)

Editing data, 31 - 34, 47 - 49
Efficiency, 74 - 75, 139, 161, 164
Equating test forms (*see* Linking test forms)
Error coefficient C_f, 135 - 140, 146, 193 - 194, 214
Estimation methods, ix - x, 15 - 20, 44 - 45
  PROX, 21 - 22, 28 - 45, 50 - 56, 60 - 62, 143, 149 - 150
  UCON, 56 - 65, 142 - 143, 148 - 150
  UFORM, 143 - 151, 214, 216
Expansion factors X and Y, 21 - 22, 30, 40 - 44, 50, 62, 148
Extending a variable, 87 - 93

Fit:
  analysis, 2 - 4, 23 - 24, 66 - 82
  computer example, 52 - 55, 58 - 59, 80 - 82
  correcting misfit, 181 - 190
  crude, 124 - 125
  diagnosing misfit, 170 - 180
  hand example, 69 - 79
  item fit, 52 - 55, 58 - 59, 77 - 79, 121 - 125
  link fit, 93 - 96, 98
  loop fit, 100
  person fit, 2 - 4, 76 - 77, 121 - 125, 165 - 180, 205 - 209
  response fit, 69 - 77, 121 - 125, 165 - 180
  ruler for fit analysis, 208 - 209
  summary of fit analysis, 79 - 80
  table for fit analysis, 73, 216
Fumbling (*see* Response pattern)

Guessing (*see* Respone pattern)

Identity line, 89, 92 - 95
Individualized testing (*see* Tailoring)
Information I_{fi}, 16 - 17, 73 - 75, 135, 161 - 164
Intensifying a variable, 87 - 94
Interval distribution of items or persons, 130 - 131, 133 - 134, 137, 139
Item:
  characteristic curve, 12 - 14, 51 - 53, 58 - 59
  difficulty calibration δ and d, 17 - 22, 25, 30, 34 - 38, 40 - 42, 54 - 55, 61 - 65
  discrimination, ix - x
    index, 52 - 55
  fit, 52 - 55, 58 - 59, 77 - 79, 121 - 125
  p-value, viii, xi - xiii, 25 - 26
  point biserial, viii, x, 26
  score s_i, 10, 18 - 22, 32 - 35
Item bank building, 98 - 118
  KCT example, 106 - 118
  chain, 99
  link, 96 - 106
  loop, 100
  network, 101 - 103
  web, 102 - 106
Item calibration quality control, 121 - 125 (*see* Item fit)

KCT (*see* Knox Cube Test)
Knox Cube Test KCT, 28 - 29
  banking KCTB, 106 - 118
  criterion referencing, 118 - 121, 206 - 207
  KCTB, 106 - 121

norm referencing, 120, 126 - 128, 198 - 200, 204, 206 - 207
response matrix, 31 - 33, 66 - 69
variable definition, 83 - 91, 119 - 120, 206 - 207

Least measurable difference LMD, 132, 135, 192 - 198
Least observable difference LOD, 132, 193, 194, 196
Least significant difference LSD, 195 - 196
Linearity, vii, 7 - 9, 15, 25, 27, 191 - 192
Linking test forms, 96 - 106
common item, 108 - 109, 112 - 118
common person, 107 - 112
Link, 96 - 98 (*see* Item bank building)
fit, 94 - 96
LOD (*see* Least observable difference)
Logistic:
distribution, ix - x
function, 15, 25, 27, 36
ogive scaling factor 1.7, 21 - 22
Logit, 16 - 17, 25, 27, 30, 34, 36, 191 - 192
Log odds (*see* Logit)
Loop, 100 (*see* Item bank building)
fit, 100
LMD (*see* Least measurable difference)
LSD (*see* Least significant difference)

Map (*see* Variable definition)
Mastery referencing (*see* Scale CHIP)
Mean square residual v, 23 - 24, 26, 53, 71 - 74, 76 - 82, 165 - 170
Measure β and b (*see* Person ability measure)
Measurement target (*see* Target of measurement)
Measuring test, 131 - 133
Misfit (*see* Fit)

Network, 101 - 103 (*see* Item bank building)
NIT (*see* Scale NIT)
Nonlinearity of test scores, 7 - 9
Norm referencing, 120 - 121, 126 - 128, 151, 198 - 200, 204, 206 - 207
Normal approximation estimation PROX, 21 - 22, 28 - 45, 50 - 56, 60 - 62, 143, 149 - 150
computer algorithm, 61 - 62
computer example, 46 - 55
hand algorithm, 21 - 22, 34, 38 - 40, 42, 44
hand example, 30 - 44
hand vs. computer, 55 - 56
PROX vs. UCON, 60 - 61
Normal distribution of items or persons, 21, 130 - 131, 133 - 134, 137, 139
Normative scaling unit (*see* Scale NIT)

Objectivity, viii - xiii, 15, 141

Person:
ability measure β and b, 17 - 22, 134 - 136, 142 - 151
PROX, 37 - 39, 43 - 44, 51, 56, 61 - 62, 143, 148 - 149
UCON, 57, 61 - 65, 142 - 143, 147 - 149
UFORM, 143 - 147, 149 - 151, 212
characteristic curve, 12 - 14
fit, 2 - 4, 76 - 77, 121 - 125, 165 - 180, 205 - 209
response x_{vi}, 9 - 14, 68 - 77, 165 - 180
score r_v, 4 - 10, 18 - 22
converting to measure b_r, 21 - 22, 27, 37 - 40, 43 - 44, 61 - 65, 142 - 151
nonlinearity, 7 - 9
relative score f_r, 132, 134, 140, 144 - 146, 149, 193 - 194
test dependence, 4 - 6
Person measure quality control, 165 - 170 (*see* Person fit)
Plodding (*see* Response pattern)
Precision of measure (*see* Standard error person measure)
Probability unit (*see* Scale CHIP)
PROX (*see* Normal approximation estimation)

Quality control, 121 - 125, 165 - 170 (*see* Fit)
Quick norms, 126 - 128

Rasch model, 9 - 27
Reliability of calibration (*see* Standard error item calibration)
Reliability of measure (*see* Standard error person measure)
Reporting forms, 205 - 209
Residual (*see* Standardized residual)
Response:
curve, 9 - 14
fit, 69 - 75, 121 - 125, 165 - 180
improbability, 71 - 74
KCT matrix, 31, 33, 66 - 69
model, 9 - 14
pattern, 2 - 4, 170 - 180
fumbling, 171, 176, 178 - 180, 188 - 190
guessing, 171, 174 - 177, 181, 185 - 187
plodding, 171, 176, 178 - 180, 188
sleeping, 171 - 177, 181 - 184
Response probability scaling unit (*see* Scale CHIP)

Sample-free item calibration, vii - xiii, 15, 20, 25 - 26
Scale, 191 - 204
additive factor α, 191 - 192
CHIP, 201 - 204
linear, vii, 7 - 9, 25, 27, 191 - 192
LMD, 132, 135, 192 - 198
logit, 16 - 17, 25, 27, 30, 34, 36, 191 - 192
NIT, 198 - 200, 204
SIT, 199 - 202, 204
spacing factor γ, 191 - 198
Score (*see* Test score)
SIT (*see* Scale SIT)
Sleeping (*see* Response pattern)
Spacing factor γ (*see* Scale spacing factor)
Standardized mean square t, 77 - 80, 165 - 169
Standardized residual z and z^2, 23 - 24, 70 - 80, 121 - 125, 165 - 180, 205 - 209
Standard error:
coefficient C_f, 135 - 140, 146, 193 - 194, 214
identify line, 89, 92 - 95
item calibration $SE(d_i)$, 21 - 22, 25 - 26, 61 - 65, 143 - 146, 192

link, 96 - 98
loop, 100
person measure SE(b_v), S_v, SEM and S, 21 - 22, 27, 61 - 65, 132 - 136, 140, 192, 194 - 198
PROX item calibration, 21 - 22
PROX person measure, 21 - 22
spacing factor, 195
Substantive scaling unit (*see* Scale SIT)

Tailoring, 151 - 164
performance, 156 - 160, 164
self, 161 - 164
status, 153 - 156, 164
Target of measurement, 129 - 131
dispersion S, 129 - 131, 133 - 134, 137 - 140
distribution D, 130 - 131, 134 - 139
location M, 130 - 131, 137 - 139
Test design, 131 - 140
distribution of items or persons, 130 - 139
height H and h, 132 - 133, 137 - 140
length L, 132 - 133, 136 - 140
operating curve, 132 - 133, 138
shape, 132 - 140
width W and w, 132 - 133, 136 - 140
Test-free person measurement, vii - xiii, 15, 20, 27, 141
Test score r, 2 - 10, 18 - 20, 27
converting to measure b_r, 142 - 164
PROX, 21 - 22, 27, 37 - 40, 43 - 44, 61 - 62, 143
UCON, 62 - 65, 142 - 143
UFORM, 143 - 151, 212
Traditional test statistics, 24 - 27

UCON (*see* Unconditional maximum likelihood estimation)
Unconditional maximum likelihood estimation UCON:
computer example, 56 - 61
computing algorithm, 62 - 65
UCON vs. PROX, 60 - 61
UFORM (*see* Uniform approximation estimation)
Uniform approximation estimation UFORM, 143 - 151, 212, 214

Validity of calibration (*see* Item fit)
Validity of measurement (*see* Person fit)
Variable definition, 1 - 4, 98 - 106
KCT, 83 - 91, 119 - 120, 206 - 207

Web, 102 - 106 (*see* Item bank building)
complete, 103 - 104
incomplete, 104 - 106

NOTATION

for Persons v = 1, N		for Items i = 1, L	
β_v	ability parameter of person v	δ_i	difficulty parameter of item i
b_v	statistic estimating β_v	d_i	statistic estimating δ_i
$SE(b_v)$	standard error of statistic b_v	$SE(d_i)$	standard error of statistic d_i
r_v	observed test score of person v	s_i	observed sample score of item i
b_r	ability estimated for score r	p_i	sample p-value of item i
n_r	number of persons with score r		
y_v	test score logit of person v	x_i	sample score logit of item i
y_r	logit of test score r		
$y.$	sample mean of person logits	$x.$	test mean of item logits
V	sample variance of person logits	U	test variance of item logits
X	person logit expansion factor to adjust for test width	Y	item logit expansion factor to adjust for sample spread

x_{vi}	response of person v to item i
$p\{x_{vi} \mid \beta_v, \delta_i\}$	probability of response x_{vi} given β_v and δ_i
π_{vi}	probability of a correct response i.e. $x_{vi} = 1$
p_{vi}	estimate of π_{vi} based on b_v and d_i
p_{ri}	estimate of π_{vi} for score r based on b_r and d_i
I_{vi}	information in x_{vi} about person v and item i
z_{vi}	standardized residual of x_{vi} from estimated expectation

v_v	mean square residual for person v	v_i	mean square residual for item i
f_v	degrees of freedom in v_v	f_i	degrees of freedom in v_i
t_v	standardized mean square v_v	t_i	standardized mean square v_i

Exceptions to this notation occur when locally convenient, particularly with "s".

NOTATION
(continued)

	for Sample of N persons		for Test of L items
M	mean person ability	H	mean item difficulty
m	estimate of M	h	estimate of H
σ	standard deviation of person ability	ω	standard deviation of item difficulty
s	estimate of σ	W	item difficulty range
		w	estimate of W

e	Napierian or natural log base e = 2.71828. . .
$e^{(\beta_v - \delta_i)}$ $\exp(\beta_v - \delta_i)$	base e raised to the exponent $(\beta_v - \delta_i)$
$\sum_j^M (y_j)$	continued sum of y_j over j = 1, M
$\prod_j^M (y_j)$	continued product y_j over j = 1, M
$\ell n(y)$	natural log of y
$E\{y\}$	expected value of y
$V\{y\}$	variance of y
1.7 $2.89 = 1.7^2$ $8.35 = 2.89^2 = 1.7^4$	coefficient which brings the logistic cumulative distribution ogive to within 0.01 of the normal cumulative distribution ogive

RASCH MODEL

For a correct response

$$x_{vi} = 1$$

$$P\{x_{vi} = 1 \mid \beta_v, \delta_i\} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)] .$$

For an incorrect response

$$x_{vi} = 0$$

$$P\{x_{vi} = 0 \mid \beta_v, \delta_i\} = 1/[1 + \exp(\beta_v - \delta_i)] .$$

For either response

$$x_{vi} = 1 \text{ or } 0$$

$$P\{x_{vi} \mid \beta_v, \delta_i\} = \exp[x_{vi}(\beta_v - \delta_i)]/[1 + \exp(\beta_v - \delta_i)] .$$