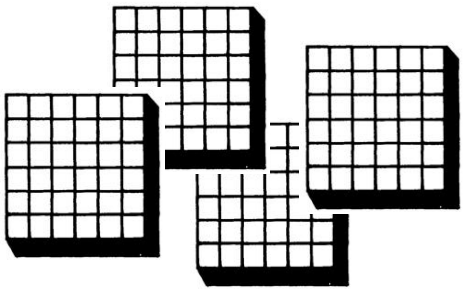
THE BASICS OF ITEM RESPONSE THEORY



FRANK B. BAKER

THE BASICS OF ITEM RESPONSE THEORY

FRANK B. BAKER
University of Wisconsin

Clearinghouse on
Assessment and Evaluation

The Basics of Item Response Theory
by Frank B. Baker
Second edition

Published by the
ERIC Clearinghouse on Assessment and Evaluation

Editors: Carol Boston, Lawrence Rudner
Design: Laura Chapman
Cover: Laura Chapman

Printed in the United States of America

ISBN 1-886047-03-0
(previously published by Heinemann, ISBN 0-435-08004-0)

This publication was set and edited with funds from the Office of Educational Research and Improvement (OERI), U.S. Department of Education, and the National Library of Education (NLE) under contract ED99CO0032. The opinions expressed in this publication do not necessarily reflect the positions or policies of OERI, the Department of Education, or NLE.

Acknowledgments

Over the past century, many people have contributed to the development of item response theory. Three persons deserve special recognition. D.N. Lawley of the University of Edinburgh published a paper in 1943 showing that many of the constructs of classical test theory could be expressed in terms of parameters of the item characteristic curve. This paper marks the beginning of item response theory as a measurement theory. The work of Dr. F.M. Lord of the Educational Testing Service has been the driving force behind both the development of the theory and its application for the past 50 years. Dr. Lord systematically defined, expanded and explored the theory as well as developed computer programs needed to put the theory into practice. This effort culminated in his classic books (with Dr. Melvin Novick, 1968; 1980) on the practical applications of item response theory. In the late 1960s, Dr. B.D. Wright of the University of Chicago recognized the importance of the measurement work by the Danish mathematician Georg Rasch. Since that time he has played a key role in bringing item response theory, the Rasch model in particular, to the attention of practitioners. Without the work of these three individuals, the level of development of item response theory would not be where it is today.

The author is indebted to Mr. T. Seavey of Heinemann Educational Books for first suggesting that I do a small book on item response theory, which resulted in the first edition of this book in 1985. This suggestion allowed me to fulfill a long-standing desire to develop an instructional software package dealing with item response theory for the then-state-of-the-art APPLE II and IBM PC computers. An upgraded version of this software has now been made available on the World Wide Web (http://ericae.net/irt).

Frank B. Baker
Madison Wisconsin

Publisher's Note

When Frank Baker wrote his classic *The Basics of Item Response Theory* in 1985, the field of educational assessment was dominated by classical test theory based on test scores. Item response theory was an upstart whose popular acceptance lagged in part because the underlying statistical calculations were quite complex. Baker's contribution was to pair a well-written introductory text on IRT with software for the then state-of-the art Apple II and IBM personal computers. The software enabled readers to experiment with concepts of the theory in eight sequential sessions.

Much has changed since 1985. IRT now powers the work of major U.S. test publishers and is used as the basis for developing the National Assessment of Educational Progress, as well as numerous state and local tests. Given its widespread acceptance, test administrators need a basic understanding of the IRT model, which this simple, well-written gem provides. We're pleased to bring it back into the public eye.

The text itself required very little updating; however, we've appended an extensive annotated list of recommended readings and Web resources. We've also updated Baker's software for the Internet (http://ericae.net/irt) to provide a new generation of readers with an interactive opportunity to explore the theory.

Lawrence A. Rudner
Director
ERIC Clearinghouse on Assessment
and Evaluation

October 2001

Table of Contents

Introduction

When this book was first published in 1985, the fields of educational measurement and psychometrics were in a transitional period. The majority of practice was based upon the classical test theory developed during the 1920s. However, a new test theory had been developing over the past forty years that was conceptually more powerful than classical test theory. Based upon items rather than test scores, the new approach was known as item response theory. While the basic concepts of item response theory were, and are, straightforward, the underlying mathematics was somewhat advanced compared to that of classical test theory. It was difficult to examine some of these concepts without performing a large number of calculations to obtain usable information. The first edition of this book was designed to provide the reader access to the basic concepts of item response theory freed of the tedious underlying calculations through an APPLE II computer program. Readers of this book may now find a new version of the program, written in Visual Basic 5.0, at (http://ericae.net/irt). Readers accustomed to sophisticated statistical and graphics packages will find it utilitarian, but nevertheless helpful in understanding various facets of the theory.

This book is organized in a building block fashion. It proceeds from the simple to the complex, with each new topic building on the preceding topics. Within each of the eight chapters, a basic concept is presented, the corresponding computer session is explained, and a set of exploratory exercises is defined. Readers are then strongly encouraged to use the computer session to explore the concept through a series of exercises. A final section of each chapter, called “Things To Notice,” lists some of the characteristics of the concept that you should have noticed and some of the conclusions you should have reached. If you do not understand the logic underlying something in this chapter, you can return to the computer session and try new variations and explorations until clarity is achieved.

When finished with the book and the computer sessions, the reader should have a good working knowledge of the fundamentals of item response theory. This book emphasizes the basics, minimizes the amount of mathematics, and does not pursue technical details that are of interest

only to the specialist. In some sense, you will be shown only "what you need to know," rather than all the glorious details of the theory. Upon completion of this book, the reader should be able to interpret test results that have been analyzed under item response theory by means of programs such as BICAL (Wright and Mead, 1976; Mislevy and Bock, 1986). In order to employ the theory in a practical setting, the reader should study more advanced books on the application of the theory. Additional print and online resources listed in the back of this new edition now supplement my original recommendations of Lord (1980), Hambleton and Swaminathan (1984), Wright and Stone (1979), or Hulin, Drasgow and Parsons (1983).

Getting Started

a. Go to ADD URL, DOWNLOAD INSTRUCTIONS.

b. A title page will be shown on the screen. Click to get the main menu.

c. Use the mouse to highlight the INTRODUCTION TO SYSTEM session and press [SELECT].

d. The following menu will appear:

USE OF ACTION BOX

YES NO RESPONSE

ENTERING NUMBERS

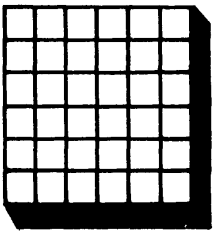
ALTERNATING DISPLAYS

RETURN TO MAIN MENU

The author recommends that you spend a few minutes with this session, even if you are an experienced computer user, to become familiar with the way the system handles the interactive procedures.

e. With the main menu on the screen, other sessions can be selected by using the mouse to highlight the session of interest, then clicking on [SELECT].

f. Once in a session, it is best to proceed through it sequentially. You may notice that the various screens you've worked through stay open on the bottom of your computer screen until you reach the closing screen for a session, which will allow you to return to the main menu. This is simply a function of the way the original software for the first edition of the book was updated.



CHAPTER 1
The Item Characteristic Curve

CHAPTER 1

The Item Characteristic Curve

In many educational and psychological measurement situations, there is an underlying variable of interest. This variable is often something that is intuitively understood, such as "intelligence." When people are described as being bright or average, the listener has some idea as to what the speaker is conveying about the object of the discussion. Similarly, one can talk about scholastic ability and its attributes, such as getting good grades, learning new material easily, relating various sources of information, and using study time effectively. In academic areas, one can use descriptive terms such as reading ability and arithmetic ability. Each of these is what psychometricians refer to as an unobservable, or latent, trait. Although such a variable is easily described, and knowledgeable persons can list its attributes, it cannot be measured directly as can height or weight, for example, since the variable is a concept rather than a physical dimension. A primary goal of educational and psychological measurement is the determination of how much of such a latent trait a person possesses. Since most of the research has dealt with variables such as scholastic, reading, mathematical, and arithmetic abilities, the generic term "ability" is used within item response theory to refer to such latent traits.

If one is going to measure how much of a latent trait a person has, it is necessary to have a scale of measurement, i.e., a ruler having a given metric. For a number of technical reasons, defining the scale of measurement, the numbers on the scale, and the amount of the trait that the numbers represent is a very difficult task. For the purposes of the first six chapters, this problem shall be solved by simply defining an arbitrary underlying ability scale. It will be assumed that, whatever the ability, it can be measured on a scale having a midpoint of zero, a unit of measurement of one, and a range from negative infinity to positive infinity. Since there is a unit of measurement and an arbitrary zero point, such a scale is referred to as existing at an interval level of measurement.

The underlying idea here is that if one could physically ascertain the ability of a person, this ruler would be used to tell how much ability a given person has, and the ability of several persons could be compared. While the theoretical range of ability is from negative infinity to positive infinity, practical considerations usually limit the range of values from, say, -3 to +3. Consequently, the discussions in the text and the computer sessions will deal only with ability values within this range. However, you should be aware that values beyond this range are possible.

The usual approach taken to measure an ability is to develop a test consisting of a number of items (questions). Each of these items measures some facet of the particular ability of interest. From a purely technical point of view, such items should be free-response items for which the examinee can write any response that seems appropriate. The person scoring the test must then decide whether the response is correct or not. When the item response is determined to be correct, the examinee receives a score of one; an incorrect answer receives a score of zero, i.e., the item is dichotomously scored. Under classical test theory, the examinee's raw test score would be the sum of the scores received on the items in the test. Under item response theory, the primary interest is in whether an examinee got each individual item correct or not, rather than in the raw test score. This is because the basic concepts of item response theory rest upon the individual items of a test rather than upon some aggregate of the item responses such as a test score.

From a practical point of view, free-response items are difficult to use in a test. In particular, they are difficult to score in a reliable manner. As a result, most tests used under item response theory consist of multiple-choice items. These are scored dichotomously: the correct answer receives a score of one, and each of the distractors yields a score of zero. Items scored dichotomously are often referred to as binary items.

A reasonable assumption is that each examinee responding to a test item possesses some amount of the underlying ability. Thus, one can consider each examinee to have a numerical value, a score, that places him or her somewhere on the ability scale. This ability score will be denoted by the Greek letter theta, θ. At each ability level, there will be a certain

probability that an examinee with that ability will give a correct answer to the item. This probability will be denoted by P(*ê*). In the case of a typical test item, this probability will be small for examinees of low ability and large for examinees of high ability. If one plotted P(*ê*) as a function of ability, the result would be a smooth S-shaped curve such as shown in Figure 1-1. The probability of correct response is near zero at the lowest levels of ability. It increases until at the highest levels of ability, the probability of correct response approaches 1. This S-shaped curve describes the relationship between the probability of correct response to an item and the ability scale. In item response theory, it is known as the item characteristic curve. Each item in a test will have its own item characteristic curve.

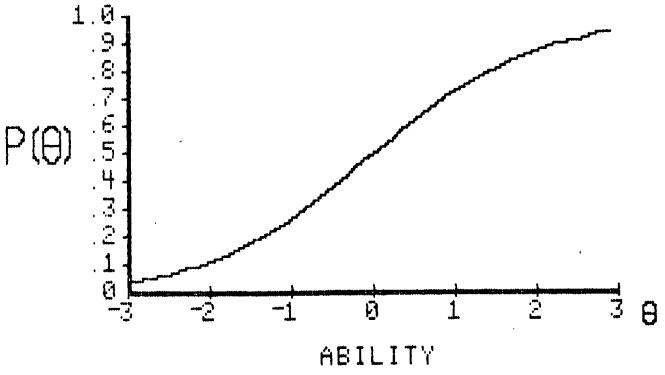


FIGURE 1-1. A typical item characteristic curve

The item characteristic curve is the basic building block of item response theory; all the other constructs of the theory depend upon this curve. Therefore, considerable attention will be devoted to this curve and its role within the theory. There are two technical properties of an item characteristic curve that are used to describe it. The first is the difficulty of the item. Under item response theory, the difficulty of an item describes where the item functions along the ability scale. For example, an easy item functions among the low-ability examinees and a hard item functions among the high-ability examinees; thus, difficulty is a location index. The second technical property is discrimination, which describes how well an item can differentiate between examinees having abilities below the item location and those having abilities above the item location. This property essentially reflects the steepness of the item

characteristic curve in its middle section. The steeper the curve, the better the item can discriminate. The flatter the curve, the less the item is able to discriminate since the probability of correct response at low ability levels is nearly the same as it is at high ability levels. Using these two descriptors, one can describe the general form of the item characteristic curve. These descriptors are also used to discuss the technical properties of an item. It should be noted that these two properties say nothing about whether the item really measures some facet of the underlying ability or not; that is a question of validity. These two properties simply describe the form of the item characteristic curve.

The idea of item difficulty as a location index will be examined first. In Figure 1-2, three item characteristic curves are presented on the same graph. All have the same level of discrimination but differ with respect to difficulty. The left-hand curve represents an easy item because the probability of correct response is high for low-ability examinees and approaches 1 for high-ability examinees. The center curve represents an item of medium difficulty because the probability of correct response is low at the lowest ability levels, around .5 in the middle of the ability scale and near 1 at the highest ability levels. The right-hand curve represents a hard item. The probability of correct response is low for most of the ability scale and increases only when the higher ability levels are reached. Even at the highest ability level shown (+3), the probability of correct response is only .8 for the most difficult item.

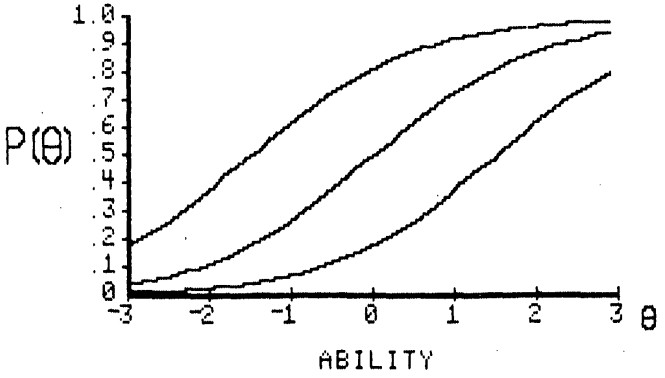


FIGURE 1-2. Three item characteristic curves with the same discrimination but different levels of difficulty

The concept of discrimination is illustrated in Figure 1-3. This figure contains three item characteristic curves having the same difficulty level but differing with respect to discrimination. The upper curve has a high level of discrimination since the curve is quite steep in the middle where the probability of correct response changes very rapidly as ability increases. Just a short distance to the left of the middle of the curve, the probability of correct response is much less than .5, and a short distance to the right the probability is much greater than .5. The middle curve represents an item with a moderate level of discrimination. The slope of this curve is much less than the previous curve and the probability of correct response changes less dramatically than the previous curve as the ability level increases. However, the probability of correct response is near zero for the lowest-ability examinees and near 1 for the highest-ability examinees. The third curve represents an item with low discrimination. The curve has a very small slope and the probability of correct response changes slowly over the full range of abilities shown. Even at low ability levels, the probability of correct response is reasonably large, and it increases only slightly when high ability levels are reached. The reader should be warned that although the figures only show a range of ability from -3 to +3, the theoretical range of ability is from negative infinity to positive infinity. Thus, all item characteristic curves of the type used here actually become asymptotic to a probability

of zero at one tail and to 1.0 at the other tail. The restricted range employed in the figures is necessary only to fit the curves on the computer screen reasonably.

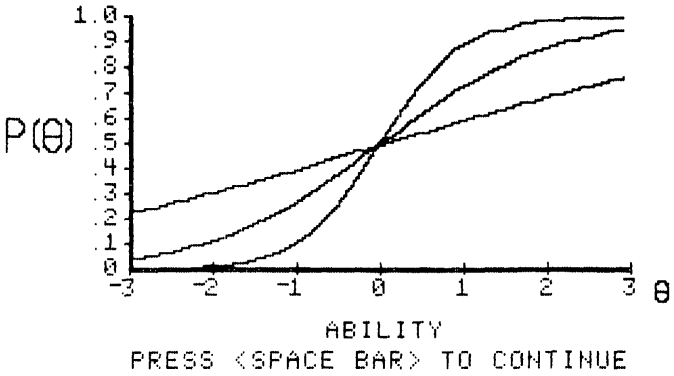


FIGURE 1-3. Three item characteristic curves with the same difficulty but with different levels of discrimination

One special case is of interest--namely, that of an item with perfect discrimination. The item characteristic curve of such an item is a vertical line at some point along the ability scale. Figure 1-4 shows such an item. To the left of the vertical line at $\grave{e} = 1.5$, the probability of correct response is zero; to the right of the line, the probability of correct response is 1. Thus, the item discriminates perfectly between examinees whose abilities are above and below an ability score of 1.5. Such items would be ideal for distinguishing between examinees with abilities just above and below 1.5. However, such an item makes no distinction among those examinees with abilities above 1.5 nor among those examinees with abilities below 1.5.

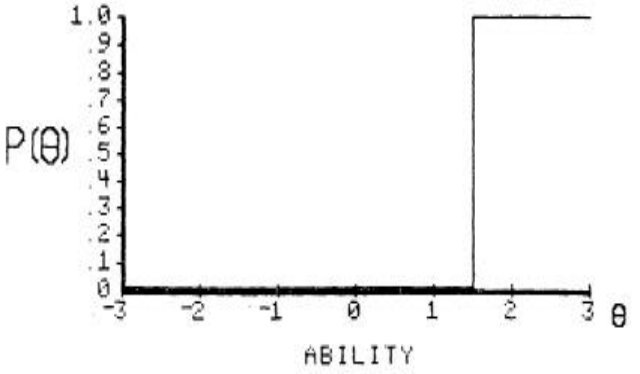


FIGURE 1-4. An item that discriminates perfectly at $\grave{e} = 1.5$

At the present point in the presentation of item response theory, the goal is to allow you to develop an intuitive understanding of the item characteristic curve and its properties. In keeping with this goal, the difficulty and discrimination of an item will be defined in verbal terms.

Difficulty will have the following levels:

very easy
easy
medium
hard
very hard

Discrimination will have the following levels:

none
low
moderate
high
perfect

These terms will be used in the computer session to specify item characteristic curves.

Computer Session for Chapter 1

The purpose of this session is to enable you to develop a sense of how

the shape of the item characteristic curve is related to item difficulty and discrimination. To accomplish this, you will be able to select verbal terms describing the difficulty and discrimination of an item. The computer will then calculate and display the corresponding item characteristic curve on the screen. You should do the exercises, then try various combinations of levels of difficulty and discrimination and relate these to the resulting curves. After a bit of such exploratory practice, you should be able to predict what the item characteristic curve will look like for a given combination of difficulty and discrimination.

Procedures for an Example Case

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight the ITEM CHARACTERISTIC CURVE session and click on [SELECT] to activate the session.

c. Read the explanatory screen and move to the next screen by clicking on [CONTINUE]. The SELECT CHARACTERISTICS screen will appear.

d. Use the left mouse button to click on medium difficulty and then click on moderate discrimination.

e. Click on [CONTINUE] to display the plot of the item characteristic curve.

f. The computer will display an item characteristic curve for an item with medium difficulty and moderate discrimination, shown in Figure 1-5.

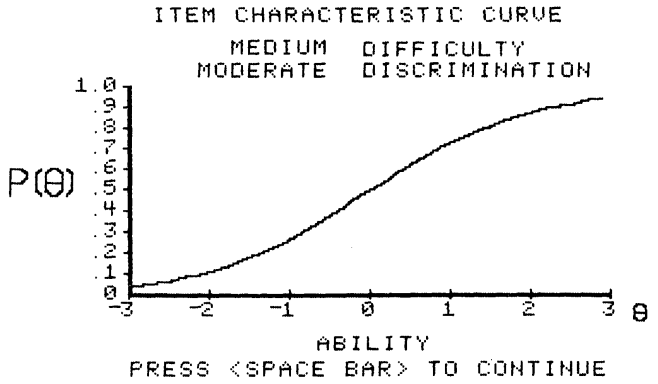


FIGURE 1-5. Example item characteristic curve

g. After you have studied the curve, click on [CONTINUE].

h. Respond to the message DO ANOTHER ITEM? by clicking on the YES button.

i. Respond to the message PLOT ON THE SAME GRAPH? by clicking on the YES button.

j. Now select easy difficulty and low discrimination and click on [CONTINUE] to see the new graph.

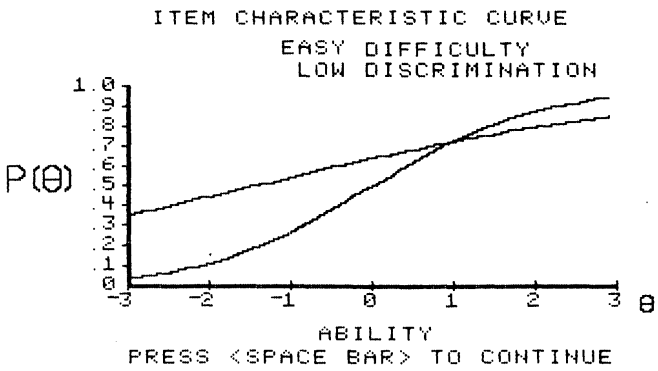


FIGURE 1-6. Two item characteristic curves

k. This second curve was overlaid on the same graph as the previous curve for comparison purposes. The new curve is rather flat and has higher probabilities of correct response in the lower range of abilities than did the previous item. This is because it was an easier item and low-ability examinees should do well on it. The low discrimination shows up in the curve having only a slight bend over the range of ability scores employed. At the high ability levels, the probability of correct response was somewhat lower than that of the previous item. This is a reflection of the lower discrimination of the new item.

l. Click on [CONTINUE] and respond to the DO ANOTHER ITEM message by clicking on the YES button.

m. To clear the graph for the next problem, respond to the message PLOT ON THE SAME GRAPH? by clicking on the NO button.

Exercises

Exercise 1

a. Use the menu to select an item with easy difficulty and high discrimination.

b. From the graph it can be seen that the probability of correct response will be rather high over most of the ability scale. The item characteristic curve will be steep in the lower part of the ability scale.

c. After you have studied the curve, respond to the message DO ANOTHER ITEM? by clicking on the YES button.

d. Respond to the message PLOT ON THE SAME GRAPH? by clicking on the NO button.

Exercise 2

a. Use the menu to select an item with hard difficulty and low discrimination.

b. From the graph it can be seen that the probability of correct response will have a low general level over most of the ability scale. The item characteristic curve will not be very steep.

c. After you have studied the curve, respond to the message DO ANOTHER ITEM? by clicking on the YES button.

d. Respond to the message PLOT ON THE SAME GRAPH? by clicking on the NO button.

Exercise 3

a. Use the menu to select an item with medium difficulty and low discrimination.

b. From the graph it can be seen that the probability of correct response will be between .2 and .8 over the range of ability shown. The item characteristic curve will be nearly linear over the range of ability employed.

c. After you have studied the curve, respond to the message DO ANOTHER ITEM? by clicking on the YES button.

d. Respond to the message PLOT ON THE SAME GRAPH? by clicking on the NO button.

Exercise 4

In this exercise, all the items will have the same difficulty but different levels of discrimination. The intent is to relate the steepness of the curves to the level of discrimination.

e. Use the menu to select an item with medium difficulty and moderate discrimination.

f. From the graph it can be seen that the probability of correct response will be small at low ability levels and large at high ability levels. The item characteristic curve will be moderately steep in the middle part of the ability scale.

g. After you have studied the curve, respond to the message DO ANOTHER ITEM? by clicking on the YES button.

h. Respond to the message PLOT ON SAME GRAPH? by clicking on the YES button.

i. Now repeat steps a through d several times using medium difficulty for each item and discrimination values of your choosing.

j. After the last item characteristic curve has been shown, clear the graph for the next problem by responding to the message PLOT ON THE SAME GRAPH? by clicking on the NO button.

Exercise 5

In this exercise, all the items will have the same level of discrimination but different difficulty levels. The intent is to relate the location of the item on the ability scale to its difficulty level.

a. Use the menu to select an item with very easy difficulty and moderate discrimination.

b. From the graph it can be seen that the probability of correct response will be reasonably large over most of the ability scale. The item characteristic curve will be moderately steep in the lower part of the ability scale.

c. After you have studied the curve, respond to the message DO ANOTHER ITEM? by clicking on the YES button.

d. Respond to the message PLOT ON SAME GRAPH? by clicking on the YES button.

e. Now repeat steps a through d several times using items with moderate discrimination and difficulty levels of your choosing.

f. After the last item characteristic curve has been shown, clear the graph for the next problem by responding to the message PLOT ON THE SAME GRAPH? by clicking on the NO button.

Exercise 6

Experiment with various combinations of difficulty of your own choosing until you are confident that you can predict the shape of the item characteristic curve corresponding to the values chosen. You may find it useful to make a rough sketch of what you think the

curve will look like before you have the computer display it on the screen.

Things To Notice

1. When the item discrimination is less than moderate, the item characteristic curve is nearly linear and appears rather flat.

2. When discrimination is greater than moderate, the item characteristic curve is S-shaped and rather steep in its middle section.

3. When the item difficulty is less than medium, most of the item characteristic curve has a probability of correct response that is greater than .5.

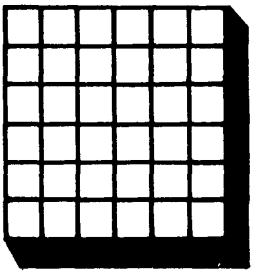
4. When the item difficulty is greater than medium, most of the item characteristic curve has a probability of correct response less than .5.

5. Regardless of the level of discrimination, item difficulty locates the item along the ability scale. Therefore item difficulty and discrimination are independent of each other.

6. When an item has no discrimination, all choices of difficulty yield the same horizontal line at a value of $P(\grave{e}) = .5$. This is because the value of the item difficulty for an item with no discrimination is undefined.

7. If you have been very observant, you may have noticed the point at which $P(\grave{e}) = .5$ corresponds to the item difficulty. When an item is easy, this value occurs at a low ability level. When an item is hard, this value corresponds to a high ability level.

CHAPTER 2



Item Characteristic Curve Models

CHAPTER 2

Item Characteristic Curve Models

In the first chapter, the properties of the item characteristic curve were defined in terms of verbal descriptors. While this is useful to obtain an intuitive understanding of item characteristic curves, it lacks the precision and rigor needed by a theory. Consequently, in this chapter the reader will be introduced to three mathematical models for the item characteristic curve. These models provide a mathematical equation for the relation of the probability of correct response to ability. Each model employs one or more parameters whose numerical values define a particular item characteristic curve. Such mathematical models are needed if one is to develop a measurement theory that can be rigorously defined and is amenable to further growth. In addition, these models and their parameters provide a vehicle for communicating information about an item's technical properties. For each of the three models, the mathematical equation will be used to compute the probability of correct response at several ability levels. Then the graph of the corresponding item characteristic curve will be shown. The goal of the chapter is to have you develop a sense of how the numerical values of the item parameters for a given model relate to the shape of the item characteristic curve.

The Logistic Function

Under item response theory, the standard mathematical model for the item characteristic curve is the cumulative form of the logistic function. It defines a family of curves having the general shape of the item characteristic curves shown in the first chapter. The logistic function was first derived in 1844 and has been widely used in the biological sciences to model the growth of plants and animals from birth to maturity. It was first used as a model for the item characteristic curve in the late 1950s and, because of its simplicity, has become the preferred model. The equation for the two-parameter logistic model is given in equation 2-1 below.

$$P(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-a(\theta-b)}} \qquad [2\text{-}1]$$

where: **e** is the constant 2.718

b is the difficulty parameter

a is the discrimination parameter[1]

L = a(è - b) is the logistic deviate (logit) and

è is an ability level.

The difficulty parameter, denoted by *b*, is defined as the point on the ability scale at which the probability of correct response to the item is .5. The theoretical range of the values of this parameter is - ∞ < = *b* < = + ∞. However, typical values have the range -3 < = *b* < = +3.

Due to the S shape of the item characteristic curve, the slope of the curve changes as a function of the ability level and reaches a maximum value when the ability level equals the item's difficulty. Because of this, the discrimination parameter does not represent the general slope of the item characteristic curve as was indicated in Chapter 1. The technical definition of the discrimination parameter is beyond the level of this book. However, a usable definition is that this parameter is proportional to the slope of the item characteristic curve at è = *b*. The actual slope at è

1. In much of the item response literature, the parameter *a* is reported as a normal ogive model value that is then multiplied by 1.70 to obtain the corresponding logistic value. This is done to make the two-parameter logistic ogive similar to the normal ogive. However, this was not done in this book because it introduces two frames of reference for interpreting the numerical values of the discrimination parameter. All item parameters in this book and the associated computer programs are interpreted in terms of the logistic function. Thus, the reported values would be divided by 1.70 to obtain the corresponding normal ogive values.

= *b* is *a*/4, but considering *a* to be the slope at *b* is an acceptable approximation that makes interpretation of the parameter easier in practice. The theoretical range of the values of this parameter is - **4**< = *a* < = + **4**, but the usual range seen in practice is -2.80 to +2.80.

Computational Example

To illustrate how the two-parameter model is used to compute the points on an item characteristic curve, consider the following example problem. The values of the item parameters are:

b = 1.0 is the item difficulty.
a = .5 is the item discrimination.

The illustrative computation is performed at the ability level *è* = -3.0.

The first term to be computed is the logistic deviate (logit), L, where:

L = *a (è - b).*

Substituting the appropriate values yields:

L = .5 (-3.0 - 1.0) = -2.0.

The next term computed is **e** (2.718) raised to the power -L. If you have a pocket calculator that can compute $\mathbf{e}^{\mathbf{x}}$ you can verify this calculation. Substituting yields:

EXP (-L) = EXP (2.0) = 7.389, where EXP represents **e**.

Now the denominator of equation 2-1 can be computed as:

1 + EXP (-L) = 1 + 7.389 = 8.389.

Finally, the value of P*(è)* is:

P*(è)* = 1/(1 + EXP (-L)) = 1/8.389 = .12.

Thus, at an ability level (T) of -3.0, the probability of responding correctly to this item is .12.

From the above, it can be seen that computing the probability of correct response at a given ability level is very easy using the logistic model. Table 2-1 shows the calculations for this item at seven ability levels evenly spaced over the range of abilities from -3 to +3. You should perform the computations at several of these ability levels to become familiar with the procedure.

TWO-PARAMETER MODEL

$$P = 1/(1 + EXP(-A(T - B)))$$
$$P = 1/(1 + EXP(- (.5) (T - (1))))$$

Ability	Logit	EXP(-L)	1 = EXP(-L)	P
-3	-2	7.389	8.389	.12
-2	-1.5	4.482	5.482	.18
-1	-1	2.718	3.718	.27
0	-.5	1.649	2.649	.38
1	0	1	2	.5
2	.5	.607	1.607	.62
3	1	.368	1.368	.73

Table 2-1. Item characteristic curve calculations under a two-parameter model, b = 1.0, a = .5

The item characteristic curve for the item of Table 2-1 is shown below. The vertical arrow corresponds to the value of the item difficulty.

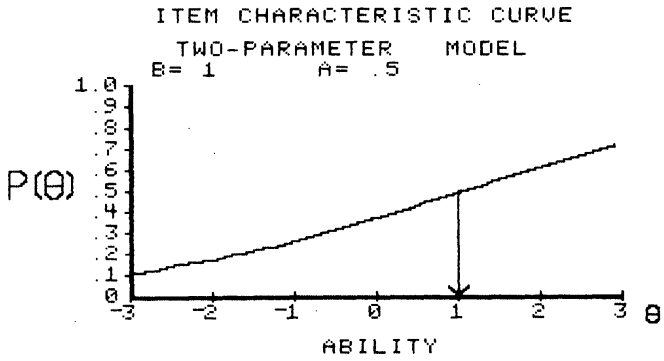


FIGURE 2-1. Item characteristic curve for a two-parameter model with b = 1.0, a = 1.5

The Rasch, or One-Parameter, Logistic Model

The next model of interest was first published by the Danish mathematician Georg Rasch in the 1960s. Rasch approached the analysis of test data from a probability theory point of view. Although he started from a very different frame of reference, the resultant item characteristic curve model was a logistic model. In Chapter 8, Rasch's approach will be explored in greater detail; our current interest is only in his item characteristic curve model. Under this model, the discrimination parameter of the two-parameter logistic model is fixed at a value of a = 1.0 for *all* items; only the difficulty parameter can take on different values. Because of this, the Rasch model is often referred to as the one-parameter logistic model.

The equation for the Rasch model is given by the following:

$$P(\theta) = \frac{1}{1 + e^{-1(\theta - b)}} \qquad [2\text{-}2]$$

where: b is the difficulty parameter and $\grave{e}$ is the ability level.

It should be noted that a discrimination parameter was used in equation 2-2, but because it always has a value of 1.0, it usually is not shown in the formula.

Computational Example

Again the illustrative computations for the model will be done for the single ability level -3.0. The value of the item difficulty parameter is:

$$b = 1.0.$$

The first term computed is the logit, L, where:

$$L = a\,(\grave{e} - b)$$

Substituting the appropriate values yields:

$$L = 1.0\ (-3.0 - 1.0) = -4.0$$

Next, the **e** to the x term is computed, giving:

$$EXP(-L) = 54.598$$

The denominator of equation 2-2 can be computed as:

$$1 + EXP(-L) = 1.0 + 54.598 = 55.598$$

Finally, the value of $P(\grave{e})$ can be obtained and is:

$$P(0) = 1/(1 + EXP(-L)) = 1/55.598 = .02$$

Thus, at an ability level of -3.0, the probability of responding correctly to this item is .02. Table 2-2 shows the calculations for seven ability levels. You should perform the computations at several other ability levels to become familiar with the model and the procedure.
The item characteristic curve corresponding to the item in Table 2-2 is

shown below.

ONE-PARAMETER MODEL

$$P = 1/(1 + EXP(-1(T - B)))$$
$$P = 1/(1 + EXP(-1(T - (1))))$$

Ability	Logit	EXP(-L)	1 = EXP(-L)	P
-3	-4	54.598	55.598	.02
-2	-3	20.086	21.086	.05
-1	-2	7.389	8.389	.12
0	-1	2.718	3.718	.27
1	0	1	2	.5
2	1	.368	1.368	.73
3	2	.135	1.135	.88

Table 2-2. Calculations for the one-parameter model, *b* = 1.0

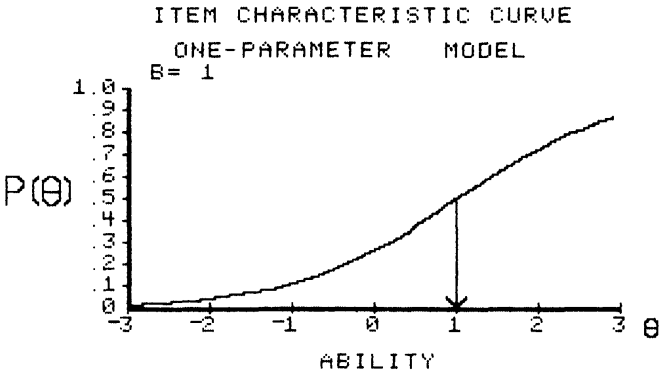


FIGURE 2-2. Item characteristic curve for a one-parameter model with *b* = 1.0

The Three-Parameter Model

One of the facts of life in testing is that examinees will get items correct by guessing. Thus, the probability of correct response includes a small component that is due to guessing. Neither of the two previous item characteristic curve models took the guessing phenomenon into consideration. Birnbaum (1968) modified the two-parameter logistic model to include a parameter that represents the contribution of guessing to the probability of correct response. Unfortunately, in so doing, some of the nice mathematical properties of the logistic function were lost. Nevertheless the resulting model has become known as the three-parameter logistic model, even though it technically is no longer a logistic model. The equation for the three-parameter model is:

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}} \qquad [2\text{-}3]$$

where: b is the difficulty parameter
a is the discrimination parameter
c is the guessing parameter and
$\grave{e}$ is the ability level

The parameter c is the probability of getting the item correct by guessing alone. It is important to note that by definition, the value of c does not vary as a function of the ability level. Thus, the lowest and highest ability examinees have the same probability of getting the item correct by guessing. The parameter c has a theoretical range of $0 < = c < = 1.0$, but in practice, values above .35 are not considered acceptable, hence the range $\grave{e}$ # c # .35 is used here.

A side effect of using the guessing parameter c is that the definition of the difficulty parameter is changed. Under the previous two models, b was the point on the ability scale at which the probability of correct response was .5. But now, the lower limit of the item characteristic curve is the value of c rather than zero. The result is that the item difficulty parameter is the point on the ability scale where:

$$P(\grave{e}) = c + (1 - c)(.5)$$

$$= (1 + c)/2$$

This probability is halfway between the value of c and 1.0. What has happened here is that the parameter c has defined a floor to the lowest value of the probability of correct response. Thus, the difficulty parameter defines the point on the ability scale where the probability of correct response is halfway between this floor and 1.0.

The discrimination parameter a can still be interpreted as being proportional to the slope of the item characteristic curve at the point $\grave{e} = b$. However, under the three-parameter model, the slope of the item characteristic curve at $\grave{e} = b$ is actually $a(1 - c)/4$.

While these changes in the definitions of parameters b and a seem slight, they are important when interpreting the results of test analyses.

Computational Example

The probability of correct response to an item under the three-parameter model will be shown for the following item parameter values:

$b = 1.5$, $a = 1.3$, $c = .2$

at an ability level of $\grave{e} = -3.0$.

The logit is:

$$L = a(\grave{e} - b) = 1.3(-3.0 - 1.5) = -5.85$$

The **e** to the x term is:

$$EXP(-L) = EXP(5.85) = 347.234$$

The next term of interest is:

$$1 + EXP(-L) = 1.0 + 347.234 = 348.234$$

and then,

$$1/(1 + EXP(-L)) = 1/348.234 = .0029$$

Up to this point, the computations are exactly the same as those for a two-parameter model with $b = 1.5$ and $a = 1.3$. But now the guessing parameter enters the picture. From equation 2-3 we have:

$P(\hat{e}) = c + (1 - c)(.0029)$ and,

c = .2 so that:

$$P(\hat{e}) = .2 + (1.0 - .2)(.0029)$$
$$= .2 + (.80)(.0029)$$
$$= .2 + (.0023)$$
$$= .2023$$

Thus, at an ability level of -3.0, the probability of responding correctly to this item is .2023. Table 2-3 shows the calculations at seven ability levels. Again, you are urged to perform the above calculations at several other ability levels to become familiar with the model and the procedures.

THREE-PARAMETER MODEL

$$P = C + (1 - C)\,(1/(1 + EXP(-A\,(T - B))))$$
$$P = .2 + (1 - .2)(1/(1 + EXP\,(-\,(1.3)\,(T - (1.5))))))$$

Ability	Logit	EXP(-L)	1 + EXP(-L)	P
-3	-5.85	347.234	348.234	.2
-2	-4.55	94.632	95.632	.21
-1	-3.25	25.79	26.79	.23
0	-1.95	7.029	8.029	.3
1	-.65	1.916	2.916	.47
2	.65	.522	1.522	.73
3	1.95	.142	1.142	.9

Table 2-3. Calculations for the three-parameter model, $b = 1.5$, $a = 1.3$, $c = .2$

The corresponding item characteristic curve is shown below.

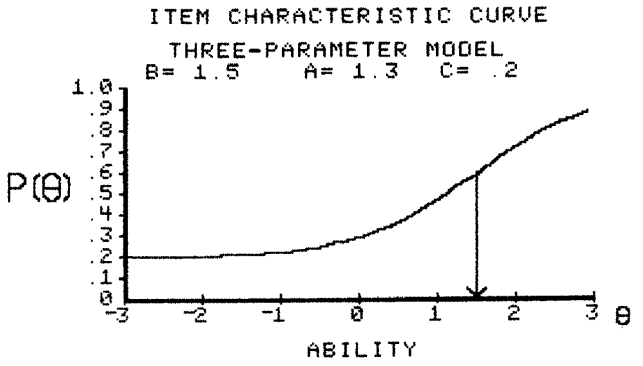


FIGURE 2-3. Item characteristic curve for a three-parameter model with b = 1.5, a = .3, c = .2

Negative Discrimination

While most test items will discriminate in a positive manner (i.e., the probability of correct response increases as the ability level increases), some items have negative discrimination. In such items, the probability of correct response decreases as the ability level increases from low to high. Figure 2-4 depicts such an item.

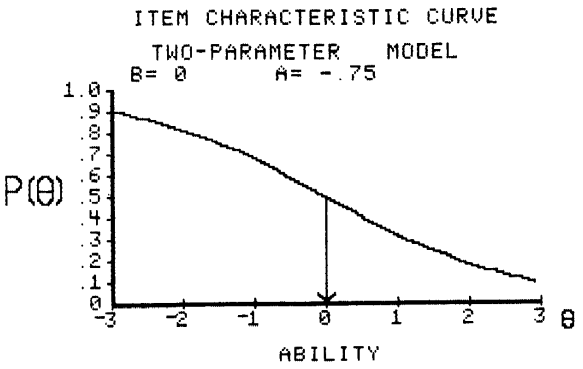


FIGURE 2-4. An item with negative discrimination under a two-parameter model with $b = 0$, $a = -.75$

Items with negative discrimination occur in two ways. First, the incorrect response to a two-choice item will always have a negative discrimination parameter if the correct response has a positive value. Second, sometimes the correct response to an item will yield a negative discrimination index. This tells you that something is wrong with the item: Either it is poorly written or there is some misinformation prevalent among the high-ability students. In any case, it is a warning that the item needs some attention. For most of the item response theory topics of interest, the value of the discrimination parameter will be positive. Figure 2-5 shows the item characteristic curves for the correct and incorrect responses to a binary item.

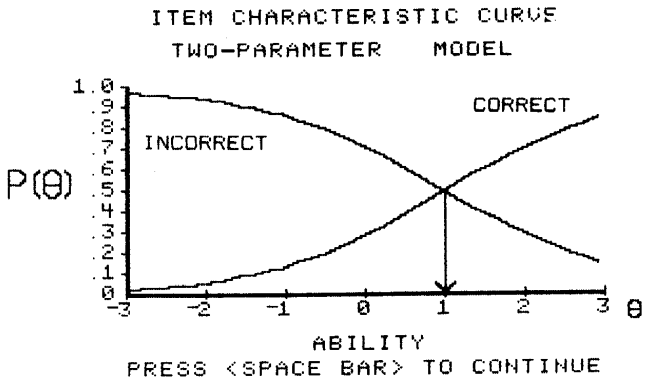


FIGURE 2-5. Item characteristic curves for the correct (b = 1.0, a = .9) and incorrect responses (b = 1.0, a = -.9) to a binary item

It should be noted that the two item characteristic curves have the same value for the difficulty parameter *(b* = 1.0) and the discrimination parameters have the same absolute value. However, they have opposite signs, with the correct response being positive and the incorrect response being negative.

Guidelines for Interpreting Item Parameter Values

In Chapter 1, verbal labels were used to describe the technical properties of an item characteristic curve. Now the curves can be described via parameters whose numerical values have intrinsic meaning. However, one needs some means of interpreting the numerical values of the item parameters and conveying this interpretation to a non-technical audience. The verbal labels used to describe an item's discrimination can be related to ranges of values of the parameter as follows:

Verbal label	Range of values
none	0
very low	.01 - .34
Low	.35 - .64
moderate	.65 - 1.34
High	1.35 - 1.69
Very high	> 1.70
Perfect	+ infinity

Table 2-4. Labels for item discrimination parameter values

These relations hold when one interprets the values of the discrimination parameter under a logistic model for the item characteristic curve. If the reader wants to interpret the discrimination parameter under a normal ogive model, divide these values by 1.7.

Establishing an equivalent table for the values of the item difficulty parameter poses some problems. The terms *easy* and *hard* used in Chapter 1 are relative terms that depend upon some frame of reference. As discussed above, the drawback of item difficulty, as defined under classical test theory, was that it was defined relative to a group of examinees. Thus, the same item could be easy for one group and hard for another group. Under item response theory, an item's difficulty is a point on the ability scale where the probability of correct response is .5 for one- and two-parameter models and $(1 + c)/2$ for a three-parameter model. Because of this, the verbal labels used in Chapter 1 have meaning only with respect to the midpoint of the ability scale. The proper way to interpret a numerical value of the item difficulty parameter is in terms of where the item functions on the ability scale. The discrimination parameter can be used to add meaning to this interpretation. The slope of the item characteristic curve is at a maximum at an ability level corresponding to the item difficulty. Thus, the item is doing its best in distinguishing between examinees in the neighborhood of this ability

level. Because of this, one can speak of the item functioning at this ability level. For example, an item whose difficulty is -1 functions among the lower ability examinees. A value of +1 denotes an item that functions among higher ability examinees. Again, the underlying concept is that the item difficulty is a location parameter.

Under a three-parameter model, the numerical value of the guessing parameter c is interpreted directly since it is a probability. For example, $c = .12$ simply means that at all ability levels, the probability of getting the item correct by guessing alone is .12.

Computer Session for Chapter 2

The purpose of this session is to enable you to develop a sense of the dependence of the shape of the item characteristic curve upon the model and the numerical values of its parameters. You will be able to set the values of the parameters under each of the three models and view the corresponding item characteristic curve on the screen. Choosing the values becomes a function of what kind of an item characteristic curve one is trying to define. Conversely, given a set of numerical values for an item characteristic curve such as provided by a test analysis, you should be able to visualize the form of the item characteristic curve from these values. Such visualization is necessary to properly interpret the technical properties of the item. After doing the exercises and a bit of exploration, you should be able to visualize the form of the item characteristic curve for any model and set of parameter values.

Procedures for an Example Case

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight ITEM PARAMETERS, then click on [CONTINUE].

c. Read the explanatory screen and click on [CONTINUE]. The SELECT ITEM CHARACTERISTIC CURVE MODEL screen will appear.

d. Use the mouse to click on TWO PARAMETER, then click on [CONTINUE]. The ITEM PARAMETER screen will appear.

e. Click on [ENTER PARAMETER VALUES], then set the values of the item parameters to a = 1.7 and b = -1.0. Select YES for VALUE(S) OK?

f. The computer will display the table of computations. Study the table for a few minutes to see the relation between the probability of correct response and the ability scores.

g. Click on [CONTINUE]. The item characteristic curve will be displayed on the screen.

h. This item functions at an ability level of -1, and the curve is quite steep at that ability level. Notice that the curve is nearly flat above an ability of 1.0. In technical terms, it is becoming asymptotic to a value of P$(\grave{e}\ )$ = 1.0.

i. Click on [CONTINUE]. The SELECT OPTIONS screen will appear.

j. At this juncture, the example case is completed. Respond to the question ANOTHER ITEM? by clicking on the YES button. Then proceed to Exercise 1.

TWO-PARAMETER MODEL

$$P = 1/(1 + EXP(-A\ (T - B)))$$
$$P = 1/(1 + EXP\ (-\ (1.7)\ (T - (-1))))$$

Ability	Logit	EXP(-L)	1 + EXP(-L)	P
-3	-3.4	29.964	30.964	.03
-2	-1.7	5.474	6.474	.15
-1	0	1	2	.5
0	1.7	.183	1.183	.85
1	3.4	.033	1.033	.97
2	5.1	6E-03	1.006	.99
3	6.8	1E-03	1.001	1

Table 2-5. Calculations for an item with *b* = -1.0, *a* = 1.7 under a two-parameter model

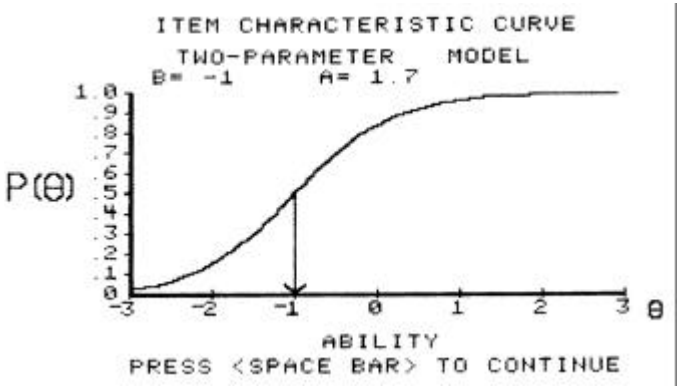


FIGURE 2-6. Item characteristic curve for *b* = -1.0, *a* = 1.7

Exercises

Exercise 1

This exercise uses the Rasch model to illustrate how the difficulty parameter locates an item along the ability scale.

a. Respond to the question SAME MODEL? by clicking on the NO button.

b. Respond to the question SHOW CALCULATIONS? by clicking on the YES button.

c. Respond to the question PLOT ON SAME GRAPH? by clicking on the NO button.

d. Respond to the question SELECTIONS OK? by clicking on the YES button. The SELECT ITEM CHARACTERISTIC CURVE MODEL screen will appear.

e. Click on RASCH, then click on [CONTINUE]. The ITEM PARAMETER VALUES screen will appear.

f. Set the value of the difficulty parameter to b = -2.0. and respond to VALUE(S) OK by clicking on the YES button.

g. The computer will display the table of computations. Study the table for a few minutes to see the relation between the probability of correct response and the ability scores.

h. Click on [CONTINUE]. The item characteristic curve will be displayed on the screen.

i. This item will function at an ability level of -2.0, and the curve will be moderately steep at that ability level. Study the plot, then click on [CONTINUE]. The SELECT OPTIONS screen will appear.

j. Next, we want to put another item characteristic curve on the same graph. Respond to the question ANOTHER ITEM? by clicking on the YES button.

k. Respond to the question SAME MODEL? by clicking on the YES button.

l. Respond to the question SHOW CALCULATIONS? by clicking on the YES button.

m. Respond to the question PLOT ON THE SAME GRAPH? by clicking on the YES button.

n. Now repeat Steps d through h, but set the item difficulty to a value of $b = 0.0$.

o. This will place a second curve on the graph.

p. Now repeat Steps i through n, using the value $b = 2.0$ for the item difficulty.

q. Now there will be three item characteristic curves on the screen. If you place a straight edge across the screen at $P(\grave{e}) = .5$, it will intersect these curves at the ability levels defined by their difficulties. In the current example, the difficulties are evenly spaced along the ability scale.

r. Click on [CONTINUE]. The SELECT OPTIONS screen will appear.

s. At this juncture, Exercise 1 is completed. Respond to the question ANOTHER ITEM? by clicking on the YES button.

Exercise 2

This exercise uses the two-parameter model to illustrate the joint effect of item difficulty and discrimination upon the shape of item characteristic curve.

a. Respond to the question SAME MODEL? by clicking on the NO button.

b. Respond to the question SHOW CALCULATIONS? by clicking on the YES button.

c. Respond to the question PLOT ON SAME GRAPH? by clicking on the NO button.

d. Respond to the question SELECTIONS OK? by clicking on the YES button. The SELECT ITEM CHARACTERISTIC CURVE MODEL screen will appear.

e. Click on TWO PARAMETER, then click on [CONTINUE]. The ITEM PARAMETER VALUES screen will appear.

f. Set the value of the item parameters to a = 1.0, b = -2.0. and respond to VALUE(S) OK? by clicking on the YES button.

g. The computer will display the table of computations. Study the table for a few minutes to see the relation between the probability of correct response and the ability scores.

h. Click on [CONTINUE]. The item characteristic curve will be displayed on the screen.

i. The item characteristic curve is located in the low-ability end of the scale and is moderately steep.

j. Next, we want to put another item characteristic curve on the same graph. Respond to the question ANOTHER ITEM? by clicking on the YES button.

k. Respond to the question SAME MODEL? by clicking on the YES button.

l. Respond to the question SHOW CALCULATIONS? by clicking on the YES button.

m. Respond to the question PLOT ON THE SAME GRAPH? by clicking on the YES button.

n. Now repeat Steps d through h, but set the item parameters to $a = 1.5$, $b = 0.0$.

o. This will place a second curve on the graph.

p. Now repeat Steps j through m, using the values $a = .5$, $b = 2.0$.

q. You should now have three item characteristic curves displayed on the same graph. It should be clear that the value of b locates the item on the ability scale and that a defines the slope. However, in the current example, the curves cross because the values of a are different for each item.

r. Click on [CONTINUE]. The SELECT OPTIONS screen will appear.

s. At this juncture, Exercise 2 is completed.

Exercise 3

This exercise illustrates the joint effect of the parameter values under the three-parameter model.

a. Respond to the question ANOTHER ITEM? by clicking on the YES button.

b. Respond to the question SAME MODEL? by clicking on the

NO button.

c. Respond to the question SHOW CALCULATIONS? by clicking on the YES button.

d. Respond to the question PLOT ON SAME GRAPH? by clicking on the NO button.

e. Respond to the question SELECTIONS OK? by clicking on the YES button. The SELECT ITEM CHARACTERISTIC CURVE MODEL screen will appear.

f. Click on THREE PARAMETER, then click on [CONTINUE]. The ITEM PARAMETER VALUES screen will appear.

g. Set the value of the item parameters to a = 1.0, b = -2.0, and c = .10, and respond to VALUE(S) OK? by clicking on the YES button.

h. The computer will display the table of computations. Study the table for a few minutes to see the relation between the probability of correct response and the ability scores.

i. Click on [CONTINUE]. The item characteristic curve will be displayed on the screen. After studying the curve, click on [CONTINUE]. The SELECT OPTIONS SCREEN will appear.

j. Respond to the question ANOTHER ITEM? by clicking on the YES button.

k. Respond to the question SAME MODEL? by clicking on the YES button.

l. Respond to the question SHOW CALCULATIONS? by clicking on the YES button.

m. Respond to the question PLOT ON SAME GRAPH? by clicking on the YES button.

n. Respond to the question SELECTIONS OK? by clicking on the YES button. The SELECT ITEM CHARACTERISTIC CURVE MODEL screen will appear.

o. Repeat steps f through h, using item parameter values of b = 0.0, a = 1.5, c = .20.

p. Repeat steps b through h, using item parameter values of b = 2.0, a = .5, c = .30.

q. At this point, you should have three item characteristic curves displayed on the graph. Again the value of b locates the items along the ability scale, but the ability level at which $P(\grave{e})$ = .5 does not correspond to the value of b but is slightly lower. Recall that under the three-parameter model, b is the point on the ability scale where the probability of correct response is $(1 + c)/2$ rather than .5. The slopes of the curves at b reflect the values of a. The lower tails of the three curves approach their values of c at the lowest levels of ability. However, this is not apparent for the curve with b = -2.0 because the values of $P(\grave{e})$ are still rather large at $\grave{e}$ = -3.0.

Exercise 4

a. For each model:

(1) Select a set of parameter values.

(2) Show at least one calculation table.

(3) Predict what the shape of the item characteristic curve will look like. It can be helpful to make a sketch of the item characteristic curve before the computer shows it on the screen.

(4) Obtain the display of the curve. (It may help to overlay a few curves to get a feeling for the relative effects of changing parameter values.)

b. Repeat this process until you know what kind of item characteristic curve will result from a set of numerical values of the item parameters under each of the models.

Things To Notice

1. Under the one-parameter model, the slope is always the same; only the location of the item changes.

2. Under the two- and three-parameter models, the value of *a* must become quite large (>1.7) before the curve is very steep.

3. Under Rasch and two-parameter models, a large positive value of *b* results in a lower tail of the curve that approaches zero. But under the three-parameter model, the lower tail approaches the value of *c*.

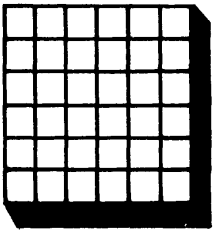
4. Under a three-parameter model, the value of *c* is not apparent when *b* < 0 and a < 1.0. However, if a wider range of values of ability were used, the lower tail would approach the value of *c*.

5. Under all models, curves with a negative value of *a* are the mirror image of curves with the same values of the remaining parameters and a positive value of *a*.

6. When *b* = -3.0, only the upper half of the item characteristic curve appears on the graph. When *b* = +3.0, only the lower half of the curve appears on the graph.

7. The slope of the item characteristic curve is the steepest at the ability level corresponding to the item difficulty. Thus, the difficulty parameter *b* locates the point on the ability scale where the item functions best.

8. Under the Rasch and two-parameter models, the item difficulty defines the point on the ability scale where the probability of correct response for persons of that ability is .5. Under a three-parameter model, the item of difficulty defines the point on the ability scale where the probability of correct response is halfway between the value of the parameter *c* and 1.0. Only when *c* = 0 are these two definitions equivalent.



CHAPTER 3
Estimating Item Parameters

CHAPTER 3

Estimating Item Parameters

Because the actual values of the parameters of the items in a test are unknown, one of the tasks performed when a test is analyzed under item response theory is to estimate these parameters. The obtained item parameter estimates then provide information as to the technical properties of the test items. To keep matters simple in the following presentation, the parameters of a single item will be estimated under the assumption that the examinees' ability scores are known. In reality, these scores are not known, but it is easier to explain how item parameter estimation is accomplished if this assumption is made.

In the case of a typical test, a sample of *M* examinees responds to the *N* items in the test. The ability scores of these examinees will be distributed over a range of ability levels on the ability scale. For present purposes, these examinees will be divided into, say, *J* groups along the scale so that all the examinees within a given group have the same ability level $\grave{e}_j$ and there will be m_j examinees within group *j*, where $j = 1, 2, 3. . . . J.$ Within a particular ability score group, r_j examinees answer the given item correctly. Thus, at an ability level of $\grave{e}_j$, the observed proportion of correct response is $p(\grave{e}_j) = r_j / m_j$, which is an estimate of the probability of correct response at that ability level. Now the value of r_j can be obtained and $p(\grave{e}_j)$ computed for each of the *j* ability levels established along the ability scale. If the observed proportions of correct response in each ability group are plotted, the result will be something like that shown in Figure 3-1.

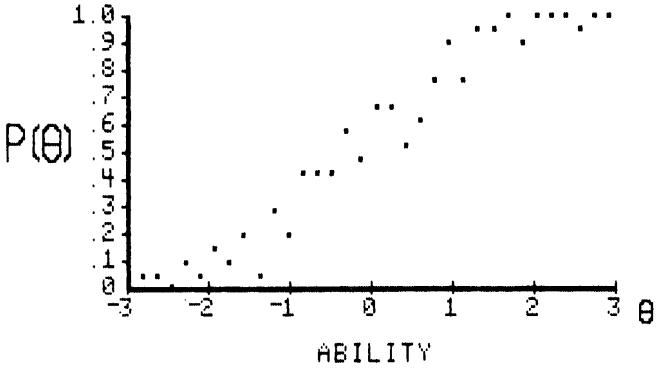


FIGURE 3-1. Observed proportion of correct response as a function of ability

The basic task now is to find the item characteristic curve that best fits the observed proportions of correct response. To do so, one must first select a model for the curve to be fitted. Although any of the three logistic models could be used, the two-parameter model will be employed here. The procedure used to fit the curve is based upon maximum likelihood estimation. Under this approach, initial values for the item parameters, such as $b = 0.0$, $a = 1.0$, are established *a priori*. Then, using these estimates, the value of $P(\grave{e}_j)$ is computed at each ability level via the equation for the item characteristic curve model. The agreement of the observed value of $p(\grave{e}_j)$ and computed value $P(\grave{e}_j)$ is determined across all ability groups. Then, adjustments to the estimated item parameters are found that result in better agreement between the item characteristic curve defined by the estimated values of the parameters and the observed proportions of correct response. This process of adjusting the estimates is continued until the adjustments get so small that little improvement in the agreement is possible. At this point, the estimation procedure is terminated and the current values of b and a are the item parameter estimates. Given these values, the equation for the item characteristic curve is used to compute the probability of correct response $P(\grave{e}_j)$ at each ability level and the item characteristic curve can be plotted. The resulting curve is the item characteristic curve

that best fits the response data for that item. Figure 3-2 shows an item characteristic curve fitted to the observed proportions of correct response shown in Figure 3-1. The estimated values of the item parameters were b = -.39 and a = 1.27.

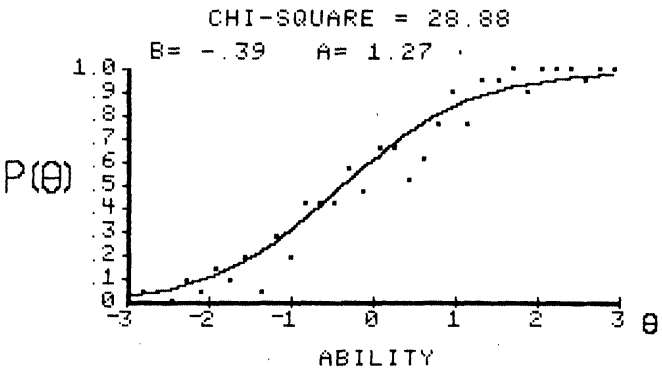


FIGURE 3-2. Item characteristic curve fitted to observed proportions of correct response

An important consideration within item response theory is whether a particular item characteristic curve model fits the item response data for an item. The agreement of the observed proportions of correct response and those yielded by the fitted item characteristic curve for an item is measured by the chi-square goodness-of-fit index. This index is defined as follows:

$$\chi^2 = \sum_{j=1}^{J} m_j \frac{[p(\theta_j) - P(\theta_j)]^2}{P(\theta_j)\ Q(\theta_j)} \qquad [3\text{-}1]$$

where: J is the number of ability groups.
$\grave{e}_j$ is the ability level of group j.
m_j is the number of examinees having ability $\grave{e}_j$.
$p(\grave{e}_j)$ is the observed proportion of correct response for group j.
$P(\grave{e}_j)$ is the probability of correct response for group j

computed from the item characteristic curve model using the item parameter estimates.

If the value of the obtained index is greater than a criterion value, the item characteristic curve specified by the values of the item parameter estimates does not fit the data. This can be caused by two things. First, the wrong item characteristic curve model may have been employed. Second, the values of the observed proportions of correct response are so widely scattered that a good fit, regardless of model, cannot be obtained. In most tests, a few items will yield large values of the chi-square index due to the second reason. However, if many items fail to yield well-fitting item characteristic curves, there may be reason to suspect that the wrong model has been employed. In such cases, re-analyzing the test under an alternative model, say the three-parameter model rather than a one-parameter model, may yield better results. In the case of the item shown in Figure 3-2, the obtained value of the chi-square index was 28.88 and the criterion value was 45.91. Thus, the two-parameter model with $b = -.39$ and $a = 1.27$ was a good fit to the observed proportions of correct response. Unfortunately, not all of the test analysis computer programs provide goodness-of-fit indices for each item in the test. For a further discussion of the model-fit issue, the reader is referred to Chapter 4 of Wright and Stone (1979).

The actual maximum likelihood estimation (MLE) procedure is rather complex mathematically and entails very laborious computations that must be performed for every item in a test. In fact, until computers became widely available, item response theory was not practical because of its heavy computational demands. For present purposes, it is not necessary to go into the details of this procedure. It is sufficient to know that the curve-fitting procedure exists, that it involves a lot of computing, and that the goodness-of-fit of the obtained item characteristic curve can be measured. Because test analysis is done by computer, the computational demands of the item parameter estimation process do not present a major problem today.

The Group Invariance of Item Parameters

One of the interesting features of item response theory is that the item parameters are not dependent upon the ability level of the examinees responding to the item. Thus, the item parameters are what is known as group invariant. This property of the theory can be described as follows. Assume two groups of examinees are drawn from the same population of examinees. The first group has a range of ability scores from -3 to -1, with a mean of -2. The second group has a range of ability scores from +1 to +3 with a mean of +2. Next, the observed proportion of correct response to a given item is computed from the item response data for every ability level within each of the two groups. Then, for the first group, the proportions of correct response are plotted as shown in Figure 3-3.

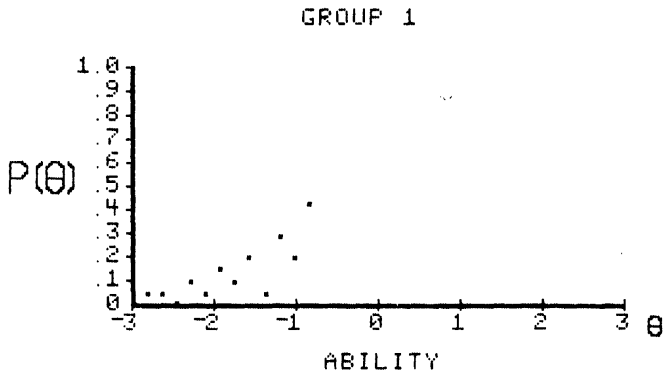


FIGURE 3-3. Observed proportions of correct response for group 1

The maximum likelihood procedure is then used to fit an item characteristic curve to the data and numerical values of the item parameter estimates, $b(1) = -.39$ and $a(1) = 1.27$, were obtained. The item characteristic curve defined by these estimates is then plotted over the range of ability encompassed by the first group. This curve is shown in Figure 3-4.

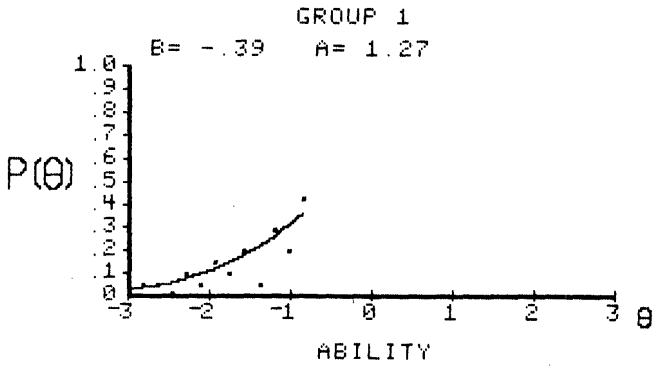


FIGURE 3-4. Item characteristic curve fitted to the group 1 data

This process is repeated for the second group. The observed proportions of correct response are shown in Figure 3-5. The fitted item characteristic curve with parameter estimates $b(2) = -.39$ and $a(2) = 1.27$ is shown in Figure 3-6.

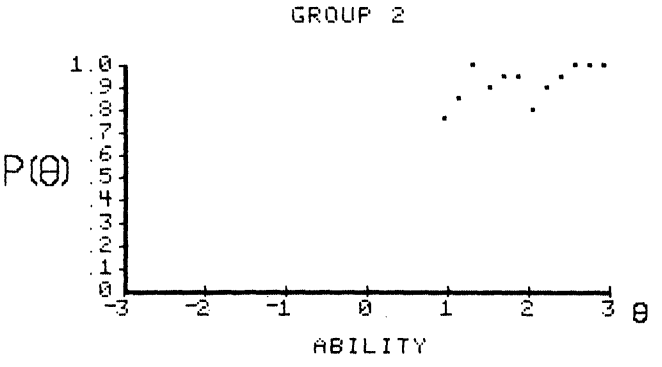


FIGURE 3-5. Observed proportions of correct response for group 2

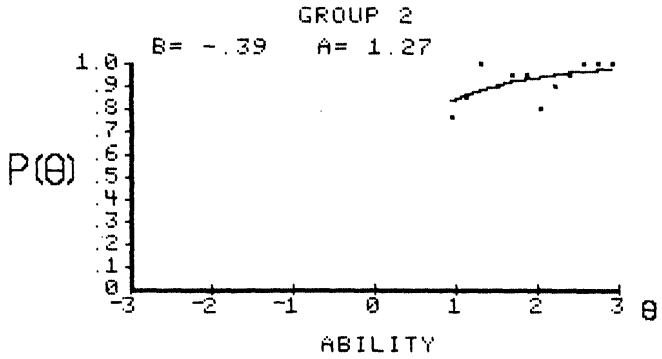


FIGURE 3-6. Item characteristic curve fitted to the group 2 data

The result of interest is that under these conditions, $b(1) = b(2)$ and $a(1) = a(2)$; i.e., the two groups yield the same values of the item parameters. Hence, the item parameters are group invariant. While this result may seem a bit unusual, its validity can be demonstrated easily by considering the process used to fit an item characteristic curve to the observed proportions of correct response. Since the first group had a low average ability (-2), the ability levels spanned by group 1 will encompass only a section of the curve, in this case, the lower left tail of the curve. Consequently, the observed proportions of correct response will range from very small to moderate values. When fitting a curve to this data, only the lower tail of the item characteristic curve is involved. For an example, see Figure 3-4. Since group 2 had a high average ability (+2), its observed proportions of correct response range from moderate to very near 1. When fitting an item characteristic curve to this data, only the upper right-hand tail of the curve is involved, as was shown in Figure 3-6. Now, since the same item was administered to both groups, the two curve-fitting processes were dealing with the same underlying item characteristic curve. Consequently, the item parameters yielded by the two analyses should be the same. Figure 3-7 integrates the two previous

diagrams into a single representation showing how the same item characteristic curve fits the two sets of proportions of correct response.

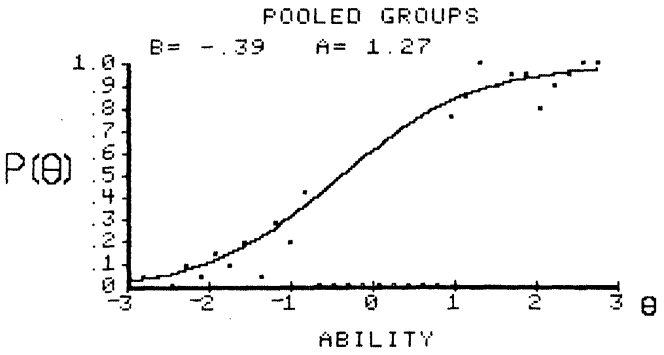


FIGURE 3-7. Item characteristic curve fitted to the pooled data, b = -.39 and a = 1.27

The group invariance of the item parameters is a very powerful feature of item response theory. It says that the values of the item parameters are a property of the item, not of the group that responded to the item. Under classical test theory, just the opposite holds. The item difficulty of classical theory is the overall proportion of correct response to an item for a group of examinees. Thus, if an item with $b = 0$ were responded to by a low-ability group, few of the examinees would get it correct. The classical item difficulty index would yield a low value, say .3, as the item difficulty for this group. If the same item were responded to by a high-ability group, most of the examinees would get it correct. The classical item difficulty index would yield a high value, say .8, indicating that the item was easy for this group. Clearly, the value of the classical item difficulty index is not group invariant. Because of this, item difficulty as defined under item response theory is easier to interpret because it has a consistent meaning that is independent of the group used to obtain its value.

WARNING: Even though the item parameters are group invariant, this does not mean that the numerical values of the item parameter estimates

yielded by the maximum likelihood estimation procedure for two groups of examinees taking the same items will always be identical. The obtained numerical values will be subject to variation due to sample size, how well-structured the data is, and the goodness-of-fit of the curve to the data. Even though the underlying item parameter values are the same for two samples, the obtained item parameter estimates will vary from sample to sample. Nevertheless, the obtained values should be "in the same ballpark." The result is that in an actual testing situation, the group-invariance principle holds but will not be apparent in the several values of the item parameter estimates obtained for the same items. In addition, the item must be used to measure the same latent trait for both groups. An item's parameters do not retain group invariance when taken out of context, i.e., when used to measure a different latent trait or with examinees from a population for which the test is inappropriate.

The group invariance of the item parameters also illustrates a basic feature of the item characteristic curve. As stated in earlier chapters, this curve is the relation between the probability of correct response to the item and the ability scale. The invariance principle reflects this since the item parameters are independent of the distribution of examinees over the ability scale. From a practical point of view, this means that the parameters of the total item characteristic curve can be estimated from any segment of the curve. The invariance principle is also one of the bases for test equating under item response theory.

Computer Session for Chapter 3

The purpose of this session is twofold. First, it serves to illustrate the fitting of item characteristic curves to the observed proportions of correct response. The computer will generate a set of response data, fit an item characteristic curve to the data under a given model, and then compute the chi-square goodness-of-fit index. This will enable you to see how well the curve-fitting procedure works for a variety of data sets and models. Second, this session shows you that the group invariance of the item parameters holds across models and over a wide range of group definitions. The session allows you to specify the range of ability encompassed by each of two groups of examinees. The computer will generate the observed proportions of correct response for each group and then fit an item characteristic curve to the data.

The values of the item parameters are also reported. Thus, you can experiment with various group definitions and observe that the group invariance holds. Example cases and exercises will be presented for both of these curve-fitting situations.

Procedures for an Example of Fitting an Item Characteristic Curve to Response Data

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight the ITEM PARAMETER ESTIMATION session and click on [SELECT].

c. Read the explanatory screen and click on [CONTINUE] to move to the SETUP screen.

d. Respond to the message SELECT NUMBER OF GROUPS by clicking on the ONE button.

e. Respond to the message SELECT ITEM CHARACTERISTIC CURVE MODEL by clicking on the TWO PARAMETER button. Then click on [CONTINUE].

f. The computer will display the observed proportion of correct response for each of 34 ability levels. The screen will be similar in appearance to Figure 3-1. The general trend of this data should suggest an item characteristic curve.

g. Click on [Plot ICC]. The computer will now fit an item characteristic curve to the observed proportions of correct response and report the values of *b* and *a*. The screen will be similar in appearance to Figure 3-2.

h. Note that the item characteristic curve defined by the estimated values of the item parameters is a good fit to the observed proportions of correct response. The obtained value of the chi-square index is less than the criterion value of 45.91.

i. After studying the graph, click on [DO ANOTHER ITEM]. The SETUP screen will appear.

Exercises

These exercises enable you to develop a sense of how well the obtained item characteristic curves fit the observed proportions of correct response. The criterion value of the chi-square index will be 45.91 for all the exercises. This criterion value actually depends upon the number of ability score intervals used and the number of parameters estimated. Thus, it will vary from situation to situation. For present purposes, it will be sufficient to use the same criterion value for all exercises.

In the next three exercises, use the ONE GROUP option.

Exercise 1

Repeat Steps c through i of the previous example several times using a Rasch model.

Exercise 2

Repeat Steps c through i several times using a two-parameter model.

Exercise 3

Repeat Steps c through i several times using a three-parameter model.

Procedures for an Example Case Illustrating Group Invariance

(Skip to Step d if you are already using this computer session.)

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight the ITEM PARAMETER ESTIMATION session and click on [SELECT].

c. Read the explanatory screen and click on [CONTINUE] to move to the SETUP screen.

d. Respond to the message SELECT NUMBER OF GROUPS by clicking on the TWO button.

e. Respond to the message SELECT ITEM CHARACTERISTIC CURVE MODEL by clicking on the TWO PARAMETER button. Then click on [CONTINUE].

f. Click on [LOWER BOUND] and set the lower bound of ability for group 1 to -3.0.

g. Click on [UPPER BOUND] and set the upper bound of ability for group 1 to -1.0.

h. Click on [LOWER BOUND] and set the lower ability bound for group 2 to +1.0.

i. Click on [UPPER BOUND] and set the upper ability bound for group 2 to +3.0.

j. Respond to the question VALUES OK? by clicking on the YES button. The INVARIANCE PRINCIPLE screen will appear.

k. Click on [DO NEXT STEP]. The plot of the observed proportions of correct response for group 1 will be shown. The screen will be similar in appearance to Figure 3-3.

l. Click on [DO NEXT STEP]. An item characteristic curve will be fitted to the data and the values of the item parameters will be reported. The screen will be similar in appearance to Figure 3-4.

m. Click on [DO NEXT STEP]. The observed proportions of correct response for group 2 will be displayed. The screen will be similar in appearance to Figure 3-5.

n. Click on [DO NEXT STEP]. The item characteristic curve will be fitted to the data and plotted for group 2, and the values of the parameters will be reported. The screen will be similar in appearance to Figure 3-6.

o. Click on [DO NEXT STEP]. The computer will now display the observed proportions of correct response for both groups on a single graph.

p. Click on [DO NEXT STEP]. An item characteristic curve will be fitted to the pooled data and the item parameters and the chi-square statistic will be reported. The numerical values will be identical to those reported for each of the two groups. The screen will be similar in appearance to Figure 3-7.

q. From this screen, it is clear that the same item characteristic curve has been fitted to both sets of data. This holds even though there was a range of ability scores (-l to +1) where there were no observed proportions of correct response to the item.

r. To do another example, click on [DO ANOTHER].

Exercises

These exercises enable you to examine the group-invariance principle under all three item characteristic curve models and for a variety of group definitions.

Exercise 1

Under a two-parameter model, set the following ability bounds:

Group 1

$LB = -2 \quad UB = +1$

Group 2

$LB = -1 \quad UB = +2$

and generate the display screens for this example.

Exercise 2

Under a one-parameter model, set the following ability bounds:

Group 1

$LB = -3 \quad UB = -1$

Group 2

$LB = +1 \quad UB = +3$

and study the resulting display screens.

Then try:

Group 1

$LB = -2 \quad UB = +1$

Group 2

LB = -1 *UB* = +2

Exercise 3

Under a three-parameter model, set the following ability bounds:

Group 1

LB = -3 *UB* = -1

Group 2

LB = +1 *UB* = +3

Then try:

Group 1

LB = -2 *UB* = +1

Group 2

LB = -1 *UB* = +2

Exercise 4

Now experiment with various combinations of overlapping and non-overlapping ability groups in conjunction with each of the three item characteristic curve models.

Things To Notice

1. Under all three models, the item characteristic curve based upon the estimated item parameters was usually a good overall fit to the observed proportions of correct response. In these exercises, this is more of a function of the manner in which the observed proportions of correct response were generated than of some intrinsic property of the item characteristic curve models. However, in most well-constructed tests, the majority of item characteristic curves specified by the item parameter estimates will fit the data. The lack of fit usually indicates that that item needs to be studied and perhaps rewritten or discarded.

2. When two groups are employed, the same item characteristic curve will be fitted, regardless of the range of ability encompassed by each group.

3. The distribution of examinees over the range of abilities for a group was not considered; only the ability levels are of interest. The number of examinees at each level does not affect the group-invariance property.

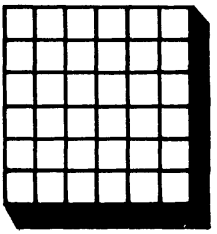
4. If two groups of examinees are separated along the ability scale and the item has positive discrimination, the low-ability group involves the lower left tail of the item characteristic curve, and the high-ability group involves the upper right tail.

5. The item parameters were group invariant whether or not the ability ranges of the two groups overlapped. Thus, overlap is not a consideration.

6. If you were brave enough to define group 1 as the high-ability group and group 2 as the low-ability group, you would have discovered that it made no difference as to which group was the high-ability group. Thus, group labeling is not a consideration.

7. The group-invariance principle holds for all three item characteristic curve models.

8. It is important to recognize that whenever item response data is used, the obtained item parameter estimates are subject to sampling variation. As a result, the same test administered to several groups of students will not yield the same numerical values for the item parameter estimates each time. However, this does not imply that the group-invariance principle is invalid. It simply means that the principle is more difficult to observe in real data.



CHAPTER 4
The Test Characteristic Curve

CHAPTER 4

The Test Characteristic Curve

Item response theory is based upon the individual items of a test, and up to this point, the chapters have dealt with the items one at a time. Now, attention will be given to dealing with all the items in a test at once. When scoring a test, the response made by an examinee to each item is dichotomously scored. A correct response is given a score of 1, and an incorrect response a score of 0; the examinee's raw test score is obtained by adding up the item scores. This raw test score will always be an integer number and will range from 0 to N, the number of items in the test. If examinees were to take the test again, assuming they did not remember how they previously answered the items, a different raw test score would be obtained. Hypothetically, an examinee could take the test a great many times and obtain a variety of test scores. One would anticipate that these scores would cluster themselves around some average value. In measurement theory, this value is known as the true score and its definition depends upon the particular measurement theory. In item response theory, the definition of a true score according to D.N. Lawley is used. The formula for a true score is given in equation 4-1 below:

$$TS_j = \sum_{i=1}^{N} P_i(\theta_j) \qquad [4\text{-}1]$$

where: TS_j is the true score for examinees with ability level $\grave{e}_j$.
i denotes an item and $P_i(\grave{e}_j)$ depends upon the particular item characteristic curve model employed.

The task at hand is to calculate the true score for those examinees having a given ability level. To illustrate this, the probability of correct response for each item in a four-item test will be calculated at an ability level of 1.0. This can be done using the formula for a two-parameter model and the procedures given in Chapter 2.

Item 1:

$P_1(1.0) = 1/(1 + EXP(-.5(1.0 - (-1.0)))) = .73$

To make the process a bit clearer, the dashed line on the figure below shows the relation between the value of $\grave{e}$ and $P_1(\grave{e})$ on the item characteristic curve.

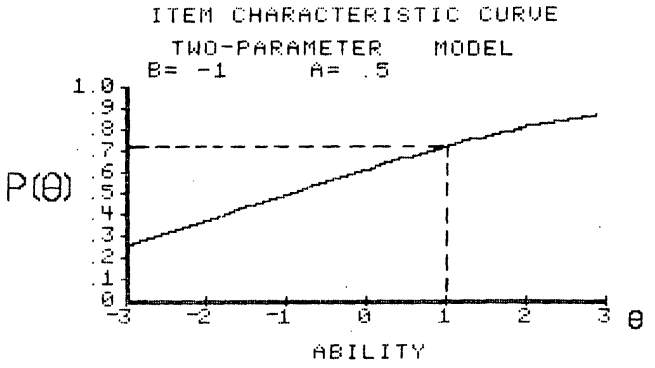


FIGURE 4-1. Item characteristic curve for item 1

Item 2:

$P_2(1.0) = 1/(1 + \exp(-1.2\ (1.0 - (.75)))) = .57$

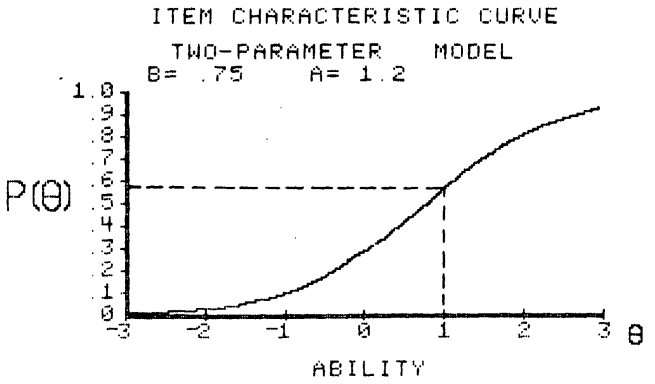


FIGURE 4-2. Item characteristic curve for item 2

Item 3:

$P_3(1.0) = 1/(1 + EXP(-.8(1.0 - (0)))) = .69$

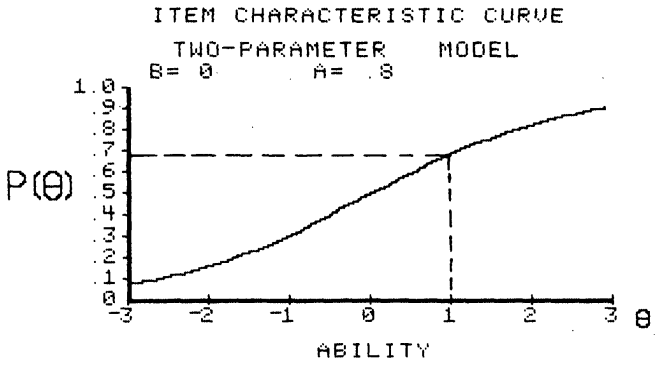


FIGURE 4-3. Item characteristic curve for item 3

Item 4:

$P_4(1.0) = 1/(1 + EXP(-1.0(1.0 - (.5)))) = .62$

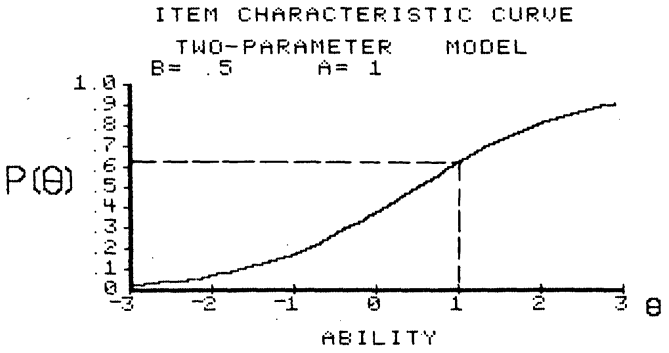


FIGURE 4-4. Item characteristic curve for item 4

Now, to get the true score at $\grave{e} = 1.0$, the probabilities of correct response are summed over the four items:

$$TS = .73 + .57 + .69 + .62 = 2.61$$

Thus, examinees having an underlying ability of 1.0 would have a true score of 2.61 on this test. This score is intuitively reasonable because at an ability score of 1.0, each of the item characteristic curves is above .5 and the sum of the probabilities would be large. While no individual examinee would actually get this score, it is the theoretical average of all the raw test scores that examinees of ability 1.0 would get on this test of four items had they taken the test a large number of times. Actual tests would contain many more items than four, but the true score would be obtained in the same manner.

The calculations performed above were for a single point on the ability scale. This true score computation could be performed for any point along the ability scale from negative infinity to positive infinity. The corresponding true scores then could be plotted as a function of ability. The vertical axis would be the true scores and would range from zero to the number of items in the test. The horizontal axis would be the ability scale. These plotted scores would form a smooth curve, the test characteristic curve. The figure below depicts a typical test characteristic curve for a test containing 10 items.

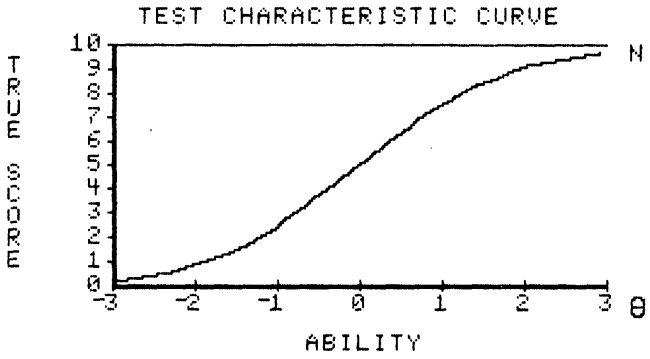


FIGURE 4-5. Test characteristic curve

The test characteristic curve is the functional relation between the true score and the ability scale. Given any ability level, the corresponding true score can be found via the test characteristic curve. For example, in Figure 4-5, draw a vertical line at an ability score of 1.0 upward until the test characteristic curve is intersected. Then, draw a horizontal line to the left until it intersects the true score scale. This line yields a true score of 7.8 for an ability score of 1.0.

When a one- or two-parameter model is used for the *N* items in a test, the left tail of the test characteristic curve approaches zero as the ability score approaches negative infinity; its upper tail approaches the number of items in the test as the ability score approaches positive infinity. The implication of this is that under these two models, a true score of zero corresponds to an ability score of negative infinity, and a true score of *N* corresponds to an ability level of positive infinity. When a three-parameter model is used for the *N* items in a test, the lower tail of the test characteristic curve approaches the sum of the guessing parameters for the test items rather than zero. This reflects the fact that under this model, very low-ability examinees can get a test score simply by guessing. The upper tail of the test characteristic curve still approaches the number of items in the test. Hence, a true score of *N* corresponds to an ability of positive infinity under all three-item characteristic curve models.

The primary role of the test characteristic curve in item response theory is to provide a means of transforming ability scores to true scores. This becomes of interest in practical situations where the user of the test may not be able to interpret an ability score. By transforming the ability score into a true score, the user is given a number that relates to the number of items in the test. This number is in a more familiar frame of reference and the user is able to interpret it. However, those familiar with item response theory, such as you, can interpret the ability score directly. The test characteristic curve also plays an important role in the procedures for equating tests.

The general form of the test characteristic curve is that of a monotonically increasing function. In some cases, it has a rather smooth S-shape similar to an item characteristic curve. In other cases, it will increase smoothly, then have a small plateau before increasing again. However, in all cases, it will be asymptotic to a value of N in the upper tail. The shape of the test characteristic curve depends upon a number of factors, including the number of items, the item characteristic curve model employed, and the values of the item parameters. Because of this, there is no explicit formula, other than equation 4-1, for the test characteristic curve as there was for the item characteristic curve. The only way one can obtain the test characteristic curve is to evaluate the probability of correct response at each ability level for all the items in the test using a given item characteristic curve model. Once these probabilities are obtained, they are summed at each ability level. Then the sums are plotted to get the test characteristic curve. It is very important to understand that the form of the test characteristic curve does not depend upon the frequency distribution of the examinees' ability scores over the ability scale. In this respect, the test characteristic curve is similar to the item characteristic curve. Both are functional relations between two scales and do not depend upon the distribution of scores over the scales.

The test characteristic curve can be interpreted in roughly the same terms as was the item characteristic curve. The ability level corresponding to the mid-true score (the number of items in the test divided by 2, i.e., $N/2$) locates the test along the ability scale. The general slope of the test characteristic curve is related to how the value of the true score depends upon the ability level. In some situations, the test characteristic curve is

nearly a straight line over much of the ability scale. In most tests, however, the test characteristic curve is nonlinear and the slope is only descriptive for a reduced range of ability levels. Since there is no explicit formula for the test characteristic curve, there are no parameters for the curve. The mid-true score defines the test difficulty in numerical terms, but the slope of the test characteristic curve is best defined in verbal terms. For most interpretive uses, these two descriptors are sufficient for discussing a test characteristic curve that has been plotted and can be visually inspected.

Computer Session for Chapter 4

This session has several purposes. The first is to show the form of the test characteristic curve and have you develop a feel for how true scores and ability are related in various tests. The second is to show the dependence of the form of the test characteristic curve upon the mix of item parameters occurring in the N items of the test. The computer session allows you to set the values of the item parameters for the N items of the test. The computer will plot the resultant test characteristic curve.

Procedures for an Example Case

This example will illustrate how to obtain a test characteristic curve for a small test.

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight the TEST CHARACTERISTIC CURVE session and click on [CONTINUE]. The TEST SPECIFICATION screen will appear.

c. Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 4$.

d. In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on TWO PARAMETER.

e. In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

f. Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

g. Click on [ENTER PARAMETERS] and set the following item parameter values:

b = -1.0, a= .50

b = .75, a= 1.2

b = 0.0, a = .8

b = .5, a = .75

h. Study the table of item parameters for a moment. If you need to change a value, click on the value and the data input box will appear, allowing you to enter a new value.

i. When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

j. Click on [CONTINUE]. The test characteristic curve shown below will appear on the screen.

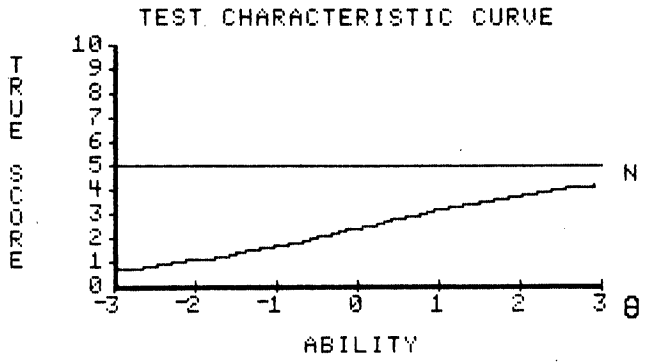


FIGURE 4-6. Test characteristic curve for the example problem

In the display, a horizontal line has been drawn at the number of items in the test. This was done to remind you that the maximum value of a true score is the number of items in the test. This technique also allows the computer program to use the same graph to show test characteristic curves based upon a different number of items.

It should be noted that because the ability range has been restricted arbitrarily to -3 to +3, the test characteristic curve may not get very close to either its theoretical upper or lower limits in the plotted curves. You should keep in mind that had the ability scale gone from negative infinity to positive infinity, the theoretical limits of the true scores would have been seen.

k. The test characteristic curve for this five-item test is very close to a straight line over the ability range from -2 to +2. Outside these values, it curves slightly. Thus, there is almost a linear relationship here between ability and true scores having a slope of about .5. The small slope reflects the low to moderate values of the item discrimination parameters. The mid-true score of 2.5 occurs at an ability level of zero, which reflects the average value of the *b*'s.

l. Click on [CONTINUE]. The SELECT OPTION FROM LIST screen will appear.

m. A list will appear. If you click on MODIFY EXISTING TEST, the ITEM PARAMETERS screen will appear, allowing you to edit any of the values. If you click on CREATE A NEW TEST, the TEST SPECIFICATION screen will appear.

Exercises

a. Using a Rasch model

Exercise 1

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 5$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on RASCH.

(4) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and set all the item difficulty parameters to a value of $b = 0.0$.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE]. The TEST CHARACTERISTIC CURVE screen will appear and the test characteristic curve will appear on the screen.

(9) The test characteristic curve in this case looks just like an item characteristic curve for an item with $b = 0$. The mid-true score occurs at an ability level of zero.

(10) Click on [CONTINUE]. The SELECT OPTION FROM LIST screen will appear. After you click on CREATE NEW TEST, the TEST SPECIFICATION screen will appear and you can do another exercise.

Exercise 2

(1) Set $N = 10$

(2) In the SELECT ITEM CHARACTERISTIC MODEL list, click on RASCH.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the difficulty parameters to:

$b = -3$, $b = -2.5$, $b = -2.0$ $b = -1.5$, $b = -1.0$, $b = -.5$, $b = 0.0$, $b = .5$, $b = 1.0$, $b = 1.5$

(5) The resulting test characteristic curve has a nearly linear section from an ability level of -3 to +1. After this point, it bends over slightly as it approaches a score of N. The mid-true score of 5 corresponds to an ability level of -.6.

(6) On the SELECT OPTION FROM LIST screen, click on CREATE NEW TEST. The TEST SPECIFICATION screen will appear and you can do the next exercise.

Exercise 3

(1) Set $N = 10$.

(2) Select a Rasch item characteristic curve model.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the item difficulty parameters to:

b = -.8, b = -.5, b = -.5, b = 0, b = 0, b = 0, b = .5, b= .5, b = .5, b = .8

(5) The test characteristic curve has a well-defined S-shape much like an item characteristic curve. Only the section near an ability level of zero is linear. The mid-true score of 2.5 corresponds to an ability score of .05.

(6) Click on [CONTINUE]. The SELECT OPTION FROM LIST screen will appear. Click on CREATE NEW TEST. The TEST SPECIFICATION screen will appear and you can do another exercise.

b. Using a two-parameter model

Exercise 1

(1) Set N = 5.

(2) Select a two-parameter item characteristic curve model.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the item parameters to:

b = 0, a = .4
b = 0, a = .8
b = 0, a = .4
b = 0, a = .8
b = 0, a = .4

(5) The test characteristic curve is nearly a straight line with a rather shallow slope reflecting the low to moderate values of a. The mid-true score of 2.5 occurs, as expected, at an ability level of 0.

Exercise 2

(1) Set $N = 5$.

(2) Select a two-parameter item characteristic curve model.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the item parameters to:

$b = 1, a = 1.6$
$b = 1, a = 1.9$
$b = 1, a = 1.6$
$b = 1, a = 1.9$
$b = 1, a = 1.6$

(5) The majority of the test characteristic curve is compressed into a rather small section of the ability scale. Up to an ability level of –1, the true score is nearly zero. Beyond an ability level of 2.5, the true score is approaching a value of N. Between these two limits, the curve has a definite S-shape, and the steep slope reflects the high values of the discrimination parameters. The mid-true score of 2.5 occurs at an ability level of 1.0. Notice how the difference in average level of discrimination in these last two problems shows in the difference of the steepness of the two curves.

Exercise 3

(1) Set $N = 5$.

(2) Select a two-parameter item characteristic curve model.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the item parameter values to:

b = -2.0,	a = .4
b = -1.5,	a = 1.7
b = -1.0,	a = .9
b = - .5,	a = 1.6
b = 0.0,	a = .8

(5) The test characteristic curve has a moderate S-shape and has a mid-true score of 2.5 at an ability level of -.8, which is not the average value of b but is close to it.

c. Using a three-parameter model

Exercise 1

(1) Set N = 5.

(2) Select a three-parameter item characteristic curve model.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the item parameters to:

b = 1.0,	a = 1.2,	c = .25
b = 1.2,	a = .9,	c = .20
b = 1.5,	a = 1.0,	c = .25
b = 1.8,	a = 1.5,	c = .20
b = 2.0,	a = .6,	c = .30

(5) The test characteristic curve has a very long lower tail that levels out just above a true score of 1.2, which is the sum of the values of the parameter c for the five items. Because of the long lower tail, there is very little change in true scores from an ability level of -3 to 0. Above an ability level of zero, the curve slopes up and begins to approach a true score of N. The mid-true score of 2.5 corresponds to an ability level of about

.5. Thus, the test functions among the high-ability examinees even though, due to guessing, low-ability examinees have true scores just above 1.2.

Exercise 2

(1) Set $N = 10$.

(2) Select a three-parameter item characteristic curve model.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the item parameters to:

$b = 2.34, a = 1.9, \quad c = .3$
$b = -1.09, a = 1.64, \quad c = .3$
$b = -1.65, a = 2.27, \quad c = .07$
$b = -.40, \quad a = .94, \quad c = .12$
$b = 2.90, \quad a = 1.83, \quad c = .16$
$b = -1.54, \quad a = 2.67, \quad c = .27$
$b = -1.52, \quad a = 2.01, \quad c = .17$
$b = -1.81, \quad a = 1.98, \quad c = .27$
$b = -.63, \quad a = .92, \quad c = .28$
$b = -2.45, \quad a = 2.54, \quad c = .07$

(5) Compared to the previous test characteristic curves, this one is quite different. The curve goes up sharply from an ability level of -3 to -1. Then it changes quite rapidly into a rather flat line that slowly approaches a value of N. The mid-true score of 5.0 is at an ability level of -1.5, indicating that the test functions among the low-ability examinees. This reflects the fact that all but two of the items had a negative value of the parameter b.

d. Exploratory exercises

(1) Set $N = 10$.

(2) Select an item characteristic curve model of your choice.

(3) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(4) Set the item parameters to values of your choice.

(5) When the test characteristic curve appears on the screen, make a sketch of it on a piece of paper.

(6) When the SELECT OPTION FROM LIST screen appears, respond to the message MODIFY EXISTING TEST? by typing YES.

(7) Change some of the values of the item parameters and then display the new test characteristic curve. Place this curve on the same graph as the previous curve.

(8) Repeat the editing process until you can predict what effect the changed item parameters will have on the form of the test characteristic curve.

(9) To explore the effect of modifications to sets of item parameters, return to Step 3 and click on COMPUTER GENERATED ITEM PARAMETER VALUES.

(10) Respond to SETTINGS OK? by selecting YES.

(11) Click on [GENERATE ITEM PARAMETER VALUES].

(12) Respond to PARAMETER VALUES OK? by selecting YES. Try to estimate the form the test characteristic curve will take, then click on [CONTINUE].

(13) After viewing the curve, press [CONTINUE] and then experiment with further modifications to the existing test.

Things To Notice

1. Relation of the true score and the ability level.

    a. Given an ability level, the corresponding true score can be found via the test characteristic curve.

    b. Given a true score, the corresponding ability level can be found via the test characteristic curve.

    c. Both the true scores and ability are continuous variables.

2. Shape of the test characteristic curve.

    a. When $N = 1$, the true score ranges from 0 to 1 and the shape of the test characteristic curve is identical to that of the item characteristic curve for the single item.

    b. The test characteristic curve does not always look like an item characteristic curve. It can have regions of varying steepness and plateaus. Such curves reflect a mixture of item parameter values having a large range of values.

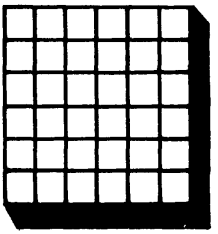
    c. The ability level at which the mid-true score ($N/2$) occurs depends primarily upon the average value of the item difficulty parameters and is an indicator of where the test functions on the ability scale.

    d. When the values of the item difficulties have a limited range, the steepness of the test characteristic curve depends primarily upon the average value of the item discrimination parameters. When the values of the item difficulties are spread widely over the ability scale, the steepness of the test characteristic curve will be reduced, even though the values of the item discriminations remain the same.

    e. Under a three-parameter model, the lower limit of the true scores is the sum of the values of the parameter c for the N items of the test.

f. The shape of the test characteristic curve depends upon the number of items in the test, the item characteristic curve model, and the mix of values of the item parameters possessed by the *N* items in the test.

3. It would be possible to construct a test characteristic curve that decreases as ability increases. To do so would require items with negative discrimination for the correct response to the items. Such a test would not be considered a good test because the higher an examinee's ability level, the lower the score expected for the examinee.



CHAPTER 5
Estimating an Examinee's Ability

CHAPTER 5

Estimating an Examinee's Ability

Under item response theory, the primary purpose for administering a test to an examinee is to locate that person on the ability scale. If such an ability measure can be obtained for each person taking the test, two goals can be achieved. First, the examinee can be evaluated in terms of how much underlying ability he or she possesses. Second, comparisons among examinees can be made for purposes of assigning grades, awarding scholarships, etc. Thus, the focus of this chapter is upon the examinees and the procedures for estimating an ability score (parameter) for an examinee.

The test used to measure an unknown latent trait will consist of N items, each of which measures some facet of the trait. In the previous chapters dealing with item parameters and their estimation, it was assumed that the ability parameter of each examinee was known. Conversely, to estimate an examinee's unknown ability parameter, it will be assumed that the numerical values of the parameters of the test items are known. A direct consequence of this assumption is that the metric of the ability scale will be the same as the metric of the known item parameters. When the test is taken, an examinee will respond to each of the N items in the test, and the responses will be dichotomously scored. The result will be a score of either a 1 or a zero for each item in the test. It is common practice to refer to the item score of 1 or 0 as the examinee's item response. Thus, the list of 1's and 0's for the N items is called the examinee's item response vector. The task at hand is to use this item response vector and the known item parameters to estimate the examinee's unknown ability parameter.

Ability Estimation Procedures

Under item response theory, maximum likelihood procedures are used to estimate an examinee's ability. As was the case for item parameter estimation, this procedure is an iterative process. It begins with some *a priori* value for the ability of the examinee and the known values of the item parameters. These are used to compute the probability of correct response to each item for that examinee. Then an adjustment to the

ability estimate is obtained that improves the agreement of the computed probabilities with the examinee's item response vector. The process is repeated until the adjustment becomes small enough that the change in the estimated ability is negligible. The result is an estimate of the examinee's ability parameter. This process is then repeated separately for each examinee taking the test. In Chapter 7, a procedure will be presented through which the ability levels of all examinees are estimated simultaneously. However, this procedure is based upon an approach that treats each examinee separately. Hence, the basic issue is how the ability of a single examinee can be estimated.

The estimation equation used is shown below:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} a_i \left[u_i - P_i(\hat{\theta}_s) \right]}{\sum_{i=1}^{N} a_i^2 \, P_i(\hat{\theta}_s) \, Q_i(\hat{\theta}_s)} \qquad [5\text{-}1]$$

where: $\hat{\theta}_s$ is the estimated ability of the examinee within iteration s

a_i is the discrimination parameter of item i, i = 1, 2, ...*N*

u_i is the response made by the examinee to item i:
$u_i = 1$ for a correct response
$u_i = 0$ for an incorrect response

$P_i(\hat{\theta}_s)$ is the probability of correct response to item i, under the given item characteristic curve model, at ability level $\hat{\theta}$ within iteration s.

$Q_i(\hat{\theta}_s) = 1 - P_i(\hat{\theta}_s)$ is the probability of incorrect response to item i, under the given item characteristic curve model, at ability level $\hat{\theta}$ within iteration s.

The equation has a rather simple explanation. Initially, the $\hat{\theta}_s$ on the right side of the equal sign is set to some arbitrary value, such as 1. The probability of correct response to each of the N items in the test is calculated at this ability level using the known item parameters in the given item characteristic curve model. Then the second term to the right

of the equal sign is evaluated. This is the adjustment term, denoted by $\Delta\hat{\theta}$. The value of $\hat{\theta}_{s+1}$ on the left side of the equal sign is obtained by adding $\Delta\hat{\theta}$ to $\hat{\theta}_s$. This value, $\hat{\theta}_{s+1}$, becomes $\hat{\theta}_s$ in the next iteration. The numerator of the adjustment term contains the essence of the procedure. Notice that $(u_i - P_i(\hat{\theta}_s))$ is the difference between the examinee's item response and the probability of correct response at an ability level of $\hat{\theta}_s$. Now, as the ability estimate gets closer to the examinee's ability, the sum of the differences between u_i and $P_i(\hat{\theta}_s)$ gets smaller. Thus, the goal is to find the ability estimate yielding values of $P_i(\hat{\theta}_s)$ for all items simultaneously that minimizes this sum. When this happens, the $\Delta\hat{\theta}$ term becomes as small as possible and the value of $\hat{\theta}_{s+1}$ will not change from iteration to iteration. This final value of $\hat{\theta}_{s+1}$ is then used as the examinee's estimated ability. The ability estimate will be in the same metric as the numerical values of the item parameters. One nice feature of equation 5-1 is that it can be used with all three item characteristic curve models, although the three-parameter model requires a slight modification.

A three-item test will be used to illustrate the ability estimation process. Under a two-parameter model, the known item parameters are:

$b = -1, \quad a = 1.0$
$b = 0, \quad a = 1.2$
$b = 1, \quad a = .8$

The examinee's item responses were:

item	response
1	1
2	0
3	1

The *a priori* estimate of the examinee's ability is set to $\hat{\theta}_s = 1.0$

First iteration:

item	*u*	*P*	*Q*	*a(u-P)*	*a*a(PQ)*
1	1	.88	.12	.119	.105
2	0	.77	.23	-.922	.255
3	1	.5	.5	.400	.160
			sum	-.403	.520

$\Delta\hat{\theta}_s = -.403/.520 = -.773$, $\hat{\theta}_{s+1} = 1.0 - .773 = .227$

Second iteration:

item	*u*	*P*	*Q*	*a(u-P)*	*a*a(PQ)*
1	1	.77	.23	.227	.175
2	0	.57	.43	-.681	.353
3	1	.35	.65	.520	.146
			sum	.066	.674

$\Delta\hat{\theta}_s = .066/.674 = .097$, $\hat{\theta}_{s+1} = .227 + .097 = .324$

Third iteration:

item	***u***	***P***	***Q***	***a(u-P)***	***a*a(PQ)***
1	1	.79	.21	.2102	.1660
2	0	.60	.40	-.7152	.3467
3	1	.37	.63	.5056	.1488
			sum	.0006	.6615

$\Delta\hat{\theta}_s = .0006/.6615 = .0009$, $\hat{\theta}_{s+1} = .324 + .0009 = .3249$

At this point, the process is terminated because the value of the adjustment (.0009) is very small. Thus, the examinee's estimated ability is .33. Unfortunately, there is no way to know the examinee's actual ability parameter. The best one can do is estimate it. However, this does not prevent us from conceptualizing such a parameter. Fortunately, one can obtain a standard error of the estimated ability that provides some indication of the precision of the estimate. The underlying principle is that an examinee, hypothetically, could take the same test a large number of times, assuming no recall of how the previous test items were answered. An ability estimate $\hat{\theta}$ would be obtained from each testing. The standard error is a measure of the variability of the values of $\hat{\theta}$ around the examinee's unknown parameter value θ. In the present case, an estimated standard error can be computed using the equation given below:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\sum_{i=1}^{N} a_i^2 \, P(\hat{\theta}) \, Q(\hat{\theta})}} \qquad [5\text{-}2]$$

It is of interest to note that the term under the square root sign is exactly the denominator of equation 5-1. As a result, the estimated standard error can be obtained as a side product of estimating the examinee's ability. In the example given above, it was:

$$SE(\hat{\theta}) = 1/\sqrt{.6615} = 1.23$$

Thus, the examinee's ability was not estimated very precisely because the standard error is very large. This is primarily due to the fact that only three items were used here and one would not expect a very good estimate. As will be shown in the next chapter, the standard error of an examinee's estimated ability plays an important role in item response theory.

There are two cases for which the maximum likelihood estimation procedure fails to yield an ability estimate. First, when an examinee answers none of the items correctly, the corresponding ability estimate is negative infinity. Second, when an examinee answers all the items in the test correctly, the corresponding ability estimate is positive infinity. In both of these cases it is impossible to obtain an ability estimate for the examinee (the computer literally cannot compute a number as big as infinity). Consequently, the computer programs used to estimate ability must protect themselves against these two conditions. When they detect either a test score of zero or a perfect test score, they will eliminate the examinee from further analysis and set the estimated ability to some symbol such as ****** to indicate what has happened.

Item Invariance of an Examinee's Ability Estimate

Another basic principle of item response theory is that the examinee's ability is invariant with respect to the items used to determine it. This

principle rests upon two conditions: first, all the items measure the same underlying latent trait; second, the values of all the item parameters are in a common metric. To illustrate this principle, assume that an examinee has an ability score of zero, which places him at the middle of the ability scale. Now, if a set of ten items having an average difficulty of -2 were administered to this examinee, the item responses could be used to estimate the examinee's ability, yielding $\hat{e}_1$ for this test. Then if a second set of ten items having an average difficulty of +1 were administered to this examinee, these item responses could be used to estimate the examinee's ability, yielding $\hat{e}_2$ for this second test. Under the item-invariance principle, $\hat{e}_1 = \hat{e}_2$; i.e., the two sets of items should yield the same ability estimate, within sampling variation, for the examinee. In addition, there is no requirement that the discrimination parameters be the same for the two sets of items. This principle is just a reflection of the fact that the item characteristic curve spans the whole ability scale. Just as any subrange of the ability scale can be used in the estimation of item parameters, the corresponding segments of several item characteristic curves can be used to estimate an examinee's ability. Items with a high average difficulty will have a point on their item characteristic curves that corresponds to the ability of interest. Similarly, items with a low average difficulty will have a point on their item characteristic curves that corresponds to the ability of interest. Consequently, either set of items can be used to estimate the ability of examinees at that point. In each set, a different part of the item characteristic curve is involved, but that is acceptable.

The practical implication of this principle is that a test located anywhere along the ability scale can be used to estimate an examinee's ability. For example, an examinee could take a test that is "easy" or a test that is "hard" and obtain, on the average, the same estimated ability. This is in sharp contrast to classical test theory, where such an examinee would get a high test score on the easy test, a low score on the hard test, and there would be no way of ascertaining the examinee's underlying ability. Under item response theory, the examinee's ability is fixed and invariant with respect to the items used to measure it. A word of warning is in order with respect to the meaning of the word "fixed." An examinee's ability is fixed only in the sense that it has a particular value in a given context. For example, if an examinee took the same test several times and it could be assumed he or she would not remember the items or the responses

from testing to testing, the examinee's ability would be fixed. However, if the examinee received remedial instruction between testings or if there were carryover effects, the examinee's underlying ability level would be different for each testing. Thus, the examinee's underlying ability level is not immutable. There are a number of applications of item response theory that depend upon an examinee's ability level changing as a function of changes in the educational context.

The item invariance of an examinee's ability and the group invariance of an item's parameters are two facets of what is referred to, generically, as the invariance principle of item response theory. This principle is the basis for a number of practical applications of the theory.

Computer Session for Chapter 5

This session has two purposes that result in apparently similar outcomes that are actually different in their conceptual basis. The first purpose is to illustrate how an examinee's estimated ability varies when the same test is taken a number of times. A test consisting of a few items will be established, the value of the examinee's ability parameter will be set, and the computer will generate the examinee's item responses. These will be used in equation 5-1 to estimate the examinee's ability. The computer will then generate a new set of item responses to these same items, and another ability estimate will be obtained. After several estimates are obtained, the mean and standard deviation of the estimates will be computed and compared to their theoretical values. The intent is to allow you to develop a sense of how ability estimates for a single examinee are distributed under repeated use of the same test.

The second purpose is to illustrate the item invariance of an examinee's ability. A small test will be established through the values of its item parameters, the examinee's ability will be set, and the computer will generate the examinee's item responses. These will be used in equation 5-1 to obtain an ability estimate for the examinee. Then a new test will be established, item responses for the same examinee generated, and another ability estimate obtained. This process will be repeated for several different tests, resulting in a set of ability estimates. If the invariance principle holds, all the estimates should be clustered around the value of the examinee's ability parameter.

Procedure for Investigating the Sampling Variability of Estimated Ability

This exercise will illustrate the sampling variability of a given examinee's estimated ability when the same test is administered several times.

n. Follow the start-up procedures described in the Introduction.

o. Use the mouse to highlight the ABILITY ESTIMATION session and click on [SELECT].

p. Read the explanatory screen and click on [CONTINUE]. The SET EXERCISE SPECIFICATIONS screen will appear.

q. After NEW EXAMINEE? click on YES.

r. Click on [NUMBER OF ABILITY ESTIMATES TO MAKE] and set the number of estimates (*K*) to 6.

s. After ADMINISTER THE SAME TEST SEVERAL TIMES? click on YES.

t. Respond to the question SPECIFICATIONS OK? by clicking on the YES button. The TEST SPECIFICATION screen will appear.

u. Click on [NUMBER OF ITEMS] and set the number of items to 5.

v. Under SELECT ITEM PARAMETER MODEL, click on TWO PARAMETER.

w. Under SELECT ITEM PARAMETER CREATION METHOD, click on USER INPUT OF ITEM PARAMETER VALUES.

x. Respond to the question SETTINGS OK? by clicking on the YES button. The ESTABLISH ITEM PARAMETER VALUES screen will appear.

y. Click on [ENTER ITEM PARAMETERS] and then set the following values:

b = -.5,	a = 1.0
b = -.25,	a = 1.5
b = 0,	a = .7
b = .25,	a = .6
b = .50,	a = 1.8

z. Study the table of item parameters for a moment. If you need to change a value, click on the value and the data input box will appear, allowing you to enter a new value.

aa. When you are satisfied with the parameter values, respond to the question PARAMETER VALUES OK by clicking on the YES button.

bb. The computer will generate and display the examinee's item responses (1 = correct, 0 = incorrect) in the table of item parameters.

cc. Click on [CONTINUE]. The ABILITY ESTIMATION RESULTS screen will appear.

dd. The computer will display the following information: the item response vector, the raw score, the estimated ability $\hat{\theta}$ and its standard error, S.E. ($\hat{\sigma}$), and the sequence number of the ability estimate.

ee. Study the results, then click on [CONTINUE]. The ABILITY ESTIMATION RESULTS screen for the next estimate will appear. Repeat K times.

ff. After the K'th ability estimate, a summary table similar to the following will appear on the screen:

TRUE ABILITY OF EXAMINEE = 1.25

ABILITY ESTIMATES

1.33 .28 .28 .69 1.25 1.86

MEAN OF ABILITY ESTIMATES = .95

STANDARD ERROR
OBSERVED = .64 THEORETICAL = .98

The mean of the ability estimates (.95 in the table above) should be reasonably close to the ability parameter value (1.25 in the table above) set by the computer for the examinee. The observed standard error of the ability estimates should approximate the theoretical value. However, with such a small number of items and replications, the results will probably deviate somewhat from their theoretical values.

a. Respond to DO ANOTHER EXERCISE? by clicking on the YES button. The EXERCISE SPECIFICATION screen will appear.

Exercises

Exercise 1

a. Respond to NEW EXAMINEE? by clicking on YES.

b. Click on [NUMBER OF ABILITY ESTIMATES TO MAKE] and set the number of estimates (*K)* to 10.

c. Respond to ADMINISTER THE SAME TEST SEVERAL TIMES? by selecting YES.

d. Respond to SPECIFICATIONS OK? by selecting YES. The TEST SPECIFICATION screen will appear.

e. Click on [NUMBER OF ITEMS] and set the number of items to 5.

f. Under SELECT ITEM PARAMETER MODEL, click on a model of your choice.

g. Under SELECT ITEM PARAMETER CREATION METHOD, click on COMPUTER GENERATED ITEM PARAMETER VALUES.

h. Respond to the question SETTINGS OK? by clicking on the YES button. The ESTABLISH ITEM PARAMETER VALUES screen will appear.

i. Click on [GENERATE ITEM PARAMETERS]. A table of item parameter values will appear. Respond to PARAMETER VALUES OK? by clicking on YES. Then click on [CONTINUE].

j. When the summary table is shown on the screen, write the value shown for the TRUE ABILITY OF EXAMINEE on a piece of paper so it can be used in the next exercise.

k. Respond to DO ANOTHER EXERCISE? by selecting YES.

Exercise 2

The intent of this exercise is to see if you can improve upon the previous estimates of the examinee's ability parameter by proper selection of the test's item parameters.

a. After NEW EXAMINEE? click on NO.

b. Click on [NUMBER OF ABILITY ESTIMATES TO MAKE] and set the number of estimates *(K)* to 10.

c. After ADMINISTER THE SAME TEST SEVERAL TIMES? click on YES.

d. Respond to SPECIFICATIONS OK? by clicking on the YES button. The TEST SPECIFICATION screen will appear.

e. Click on [NUMBER OF ITEMS] and set the number of items to 5.

f. Under SELECT ITEM PARAMETER MODEL, click on the same model as used in Exercise 1.

g. Under SELECT ITEM PARAMETER CREATION METHOD, click on USER INPUT OF ITEM PARAMETER VALUES.

h. Since you know the value of the examinee's ability parameter, choose values of the item difficulty parameters that are close to this value and use large values of the discrimination parameters.

i. Follow Steps n through s of the example case.

j. The summary table will be shown on the screen. If you chose the parameter values wisely, the mean of the ability estimates should have been close to the value of the examinee's ability parameter. The observed standard error should have also approximated the theoretical value. If such was not the case, think about some reasons for the lack of a match. You need to keep in mind that the obtained results are subject to considerable sampling variability due to the small numbers of items being used (increasing N to 10 will help) and the small number of replications used.

Exercise 3

Experiment with different types of models and item parameter values to see if you can determine what influences the distribution of the estimated abilities.

Procedures for Investigating the Item Invariance of an Examinee's Ability

In this example, a given examinee will be administered a number of different tests. The intent is to illustrate that the estimated abilities should cluster about the value of the examinee's ability parameter.

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight on the ABILITY ESTIMATION session and click on [SELECT].

c. Read the explanatory screen and click on [CONTINUE]. The SET EXERCISE SPECIFICATIONS screen will appear.

d. After NEW EXAMINEE?, click on YES.

e. Click on [NUMBER OF ABILITY ESTIMATES TO MAKE] and set the number of estimates (*K*) to 10.

f. After ADMINISTER SAME TEST SEVERAL TIMES? click on NO.

g. Respond to question SPECIFICATIONS OK? by clicking on the YES button. The TEST SPECIFICATION screen will appear.

h. Click on [NUMBER OF ITEMS] and set the number of items to 5.

i. Under SELECT ITEM PARAMETER MODEL, click on RASCH.

j. Under SELECT ITEM PARAMETER CREATION METHOD, click on COMPUTER GENERATION OF ITEM PARAMETER VALUES.

k. Respond to the question SETTINGS OK? by clicking on the YES button. The ESTABLISH ITEM PARAMETER VALUES screen will appear.

l. Click on [ENTER ITEM PARAMETERS]. The computer will generate the random item difficulty values for the 5 items.

m. Study the table of item parameters for a moment and take note of their average value and the distribution of the values.

n. When you are satisfied with the parameter values, respond to the question PARAMETERS OK? by clicking on the YES button. Then click on [CONTINUE].

o. The computer will generate and display the examinee's item responses in the table of item parameters, using 1 for a correct response and 0 for an incorrect one.

p. Click on [CONTINUE]. The ABILITY ESTIMATION RESULTS screen will appear.

q. Study the results, then click on [CONTINUE]. Repeat Steps l through q *K* times.

r. The computer will display the following information: the item response vector, the raw score, the estimated ability $\hat{\theta}$ and its standard error, S.E.$(\hat{\theta})$, and the sequence number of the ability estimate.

s. After the *K*'th ability estimate, a summary table similar to the following will appear on the screen:

ABILITY OF EXAMINEE = .50

ABILITY ESTIMATES

.11 .22 .33 .44 .55 .66 .77 .88 .99
.00

MEAN OF ABILITY ESTIMATES = .50

STANDARD ERROR
OBSERVED = 1.23 THEORETICAL = 1.35

The average value of the estimates should be close to the value of the examinee's ability parameter. There should not be a large amount of scatter in the estimates. Again, due to the small number of items and estimates used, the item invariance of the ability estimate may not be readily apparent.

t. Respond to DO ANOTHER EXERCISE? by selecting YES. The EXERCISE SPECIFICATION screen will appear.

Exercises

Exercise 1

The intent of this exercise is to enable you to experiment with the item sets used to illustrate the item invariance of the ability estimates. Rather than letting the computer set the values of the item parameters, you can choose your own values.

a. After NEW EXAMINEE?, click on YES.

b. Click on [NUMBER OF ABILITY ESTIMATES TO MAKE] and set the number of estimates (*K*) to 4.

c. After ADMINISTER SAME TEST SEVERAL TIMES?, click on NO.

d. Respond to SPECIFICATIONS OK? by clicking on YES. The TEST SPECIFICATION screen will appear.

e. Click on [NUMBER OF ITEMS] and set the number of items to 5.

f. Under SELECT ITEM PARAMETER MODEL, click on the model of your choice.

g. Under SELECT ITEM PARAMETER CREATION METHOD, click on USER INPUT OF ITEM PARAMETER VALUES.

h. Respond to the question SETTINGS OK? by clicking on the YES button. The ESTABLISH ITEM PARAMETER VALUES screen will appear.

i. Click on [ENTER ITEM PARAMETERS] and enter item parameter values for the 5 items.

j. When you are satisfied with the parameter values, respond to the question PARAMETER VALUES OK? by clicking on the YES button.

k. Click on [CONTINUE]. The ABILITY ESTIMATION RESULTS screen will appear.

l. After each estimate has been shown, enter a new set of values for the item parameters. When doing so, experiment with their average values and range of values.

m. After the *K*'th estimate has been shown, the summary table will appear and you will be able to see how well you have done.

Exercise 2

To make things easy for you, let the computer generate the sets of item parameter values. Repeat the procedures for exercise 1, but at Step h, respond to DO YOU WANT TO SET THE VALUES OF THE ITEM PARAMETERS? by selecting NO. Now the computer will do the tedious job of setting the parameters.

Things To Notice

1. Distribution of estimated ability.

    a. The average value of the estimates is reasonably close to the value of the ability parameter for the examinee set by the computer program.

    b. When the item difficulties are at or near the examinee's ability parameter value, the mean of the estimated abilities will be close to that ability value.

    c. The standard error of the estimates can be quite large when the items are not located near the ability of the examinee. However, the theoretical values of the standard errors are also quite large, and the obtained standard errors approximate these values.

    d. When the values of the item discrimination indices are large, the standard error of the ability estimates is small. When the item discrimination indices are small, the standard error of the ability estimates is large.

    e. The optimum set of items for estimating an examinee's ability would have all its item difficulties equal to the examinee's ability parameter and have items with large values for the item discrimination indices.

2. Item invariance of the examinee's ability.

    a. The different sets of items yielded values of estimated ability that were near the examinee's actual ability level.

    b. The mean value of these estimates generally was a close approximation of the examinee's ability parameter. If one used many tests, each having a large number of items, the mean estimated ability would equal the examinee's ability parameter. In addition, these estimates would be very tightly clustered

around the parameter value. In such a situation, it would be very clear that the item invariance principle holds.

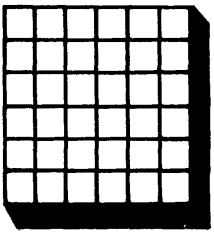
3. OveraIl observation.

   This session has dealt with two facets of estimating an examinee's ability that are conceptually distinct but look similar in certain respects. The first set of examples focused upon the variability of the ability estimates about the value of the examinee's ability parameter. This will serve as the basis for the next chapter, which deals with how well a test estimates ability over the whole ability scale. The second set of exercises focused upon the item invariance of an examinee's estimated ability. This will serve as part of the basis for Chapter 7 dealing with test calibration.

   The reader should keep in mind that an ability estimate is just another type of test score, but it is interpreted within the context of item response theory. Consequently, such ability estimates can be used to compute summary statistics for groups of examinees and other indices of interest.

4. A final comment.

   In Chapter 1, the concept of a latent trait was introduced. An integral part of item response theory is that an examinee can be positioned on the scale representing this latent trait. Thus, in theory, each examinee has an ability score (parameter value) that locates that person on the scale. However, in the real world we cannot obtain the value of the examinee's ability parameter. The best we can do is obtain an estimate of it. In the computer session for this chapter it was assumed that we could generate the value of an examinee's ability parameter. This assumption enabled the program to generate the item response vectors used to obtain the ability estimates and hence illustrate the theory.



CHAPTER 6
The Information Function

CHAPTER 6

The Information Function

When you speak of having information, it implies that you know something about a particular object or topic. In statistics and psychometrics, the term *information* conveys a similar, but somewhat more technical, meaning. The statistical meaning of information is credited to Sir R.A. Fisher, who defined information as the reciprocal of the precision with which a parameter could be estimated. Thus, if you could estimate a parameter with precision, you would know more about the value of the parameter than if you had estimated it with less precision. Statistically, the precision with which a parameter is estimated is measured by the variability of the estimates around the value of the parameter. Hence, a measure of precision is the variance of the estimators, which is denoted by σ^2. The amount of information, denoted by I, is given by the formula:

$$I = \frac{1}{\sigma^2} \qquad [6\text{-}1]$$

In item response theory, our interest is in estimating the value of the ability parameter for an examinee. The ability parameter is denoted by θ, and $\hat{\theta}$ is an estimator of θ. In the previous chapter, the standard deviation of the ability estimates about the examinee's ability parameter was computed. If this term is squared, it becomes a variance and is a measure of the precision with which a given ability level can be estimated. From equation 6-1, the amount of information at a given ability level is the reciprocal of this variance. If the amount of information is large, it means that an examinee whose true ability is at that level can be estimated with precision; i.e., all the estimates will be reasonably close to the true value. If the amount of information is small, it means that the ability cannot be estimated with precision and the estimates will be widely scattered about the true ability. Using the appropriate formula, the amount of information can be computed for each ability level on the ability scale from negative infinity to positive infinity. Because ability is a continuous variable, information will also be a continuous variable. If the amount of

information is plotted against ability, the result is a graph of the information function such as that shown below.

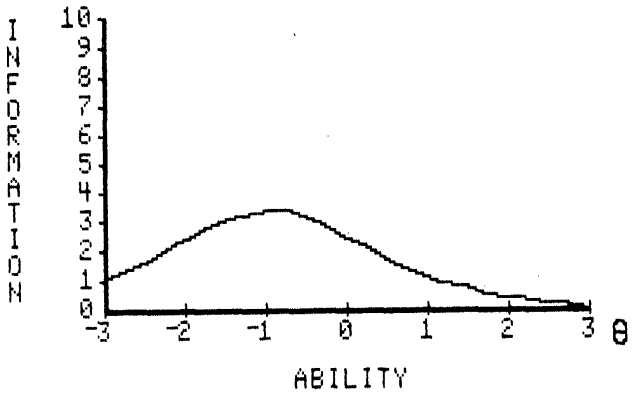


FIGURE 6-1. An information function

Inspection of Figure 6-1 shows that the amount of information has a maximum at an ability level of -1.0 and is about 3 for the ability range of $-2 <= \grave{e} <= \grave{e}$. Within this range, ability is estimated with some precision. Outside this range, the amount of information decreases rapidly, and the corresponding ability levels are not estimated very well. Thus, the information function tells us how well each ability level is being estimated. It is important for the reader to recognize that the information function does not depend upon the distribution of examinees over the ability scale. In this regard, it is like the item characteristic curve and the test characteristic curve. In a general-purpose test, the ideal information function would be a horizontal line at some large value of I and all ability levels would be estimated with the same precision. Unfortunately, such an information function is hard to achieve. The typical information function looks somewhat like that shown in Figure 6-1, and different ability levels are estimated with differing degrees of precision. This becomes of considerable importance to both the test constructor and the test consumer since it means that the precision with which an examinee's ability is estimated depends upon where the examinee's ability is located on the ability scale.

Item Information Function

Since it depends upon the individual items composing a test, item response theory is what is known as an itemized theory. Under the theory, each item of the test measures the underlying latent trait. As a result, the amount of information, based upon a single item, can be computed at any ability level and is denoted by $I_i(\theta)$, where *i* indexes the item. Because only a single item is involved, the amount of information at any point on the ability scale is going to be rather small. If the amount of item information is plotted against ability, the result is a graph of the item information function such as that shown below.

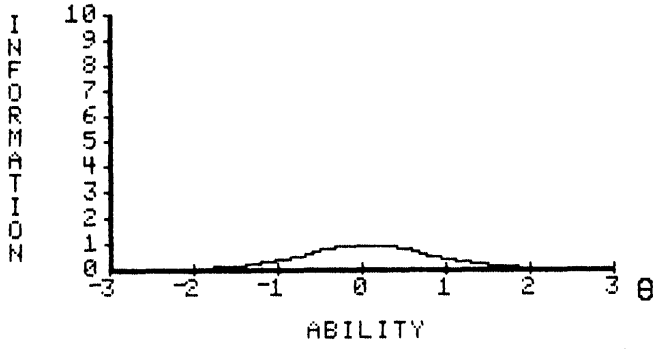


FIGURE 6-2. An item information function

An item measures ability with greatest precision at the ability level corresponding to the item's difficulty parameter. The amount of item information decreases as the ability level departs from the item difficulty and approaches zero at the extremes of the ability scale.

Test Information Function

Since a test is used to estimate the ability of an examinee, the amount of information yielded by the test at any ability level can also be obtained. A test is a set of items; therefore, the test information at a given ability level is simply the sum of the item informations at that level. Consequently, the test information function is defined as:

$$I(q) = \sum_{i=1}^{N} I_i(q) \qquad [6\text{-}2]$$

where: $I(\theta)$ is the amount of test information at an ability level of θ,
$I_i(\theta)$ is the amount of information for item i at ability level θ,
N is the number of items in the test.

The general level of the test information function will be much higher than that for a single item information function. Thus, a test measures ability more precisely than does a single item. An important feature of the definition of test information given in equation 6-2 is that the more items in the test, the greater the amount of information. Thus, in general, longer tests will measure an examinee's ability with greater precision than will shorter tests. Plotting the amount of test information against ability yields a graph of the test information function such as that shown below for a ten-item test.

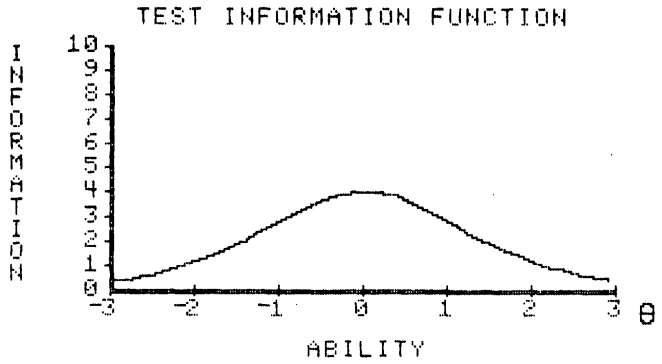


FIGURE 6-3. A test information function

The maximum value of the test information function in Figure 6-3 is modest and, in this example, the amount of information decreases rather steadily as the ability level differs from that corresponding to the maximum. Thus, ability is estimated with some precision near the center of the ability scale. However, as the ability level approaches the extremes of the scale, the amount of test information decreases significantly.

The test information function is an extremely useful feature of item response theory. It basically tells you how well the test is doing in estimating ability over the whole range of ability scores. While the ideal test information function often may be a horizontal line, it may not be the best for a specific purpose. For example, if you were interested in constructing a test to award scholarships, this ideal might not be optimal. In this situation, you would like to measure ability with considerable precision at ability levels near the ability used to separate those who will receive the scholarship from those who do not. The best test information function in this case would have a peak at the cutoff score. Other specialized uses of tests could require other forms of the test information function.

While an information function can be obtained for each item in a test, this is rarely done. The amount of information yielded by each item is rather small, and we typically do not attempt to estimate an examinee's ability with a single item. Consequently, the amount of test information at an ability level and the test information function are of primary interest. Since the test information is obtained by summing the item informations at a given ability level, the amount of information is defined at the item level. The mathematical definition of the amount of item information depends upon the particular item characteristic curve model employed. Therefore, it is necessary to examine these definitions under each model.

Definition of Item Information

Two-Parameter Item Characteristic Curve Model

Under a two-parameter model, the item information function is defined as:

$$I_i(\grave{e}) = a_i^2 P_i(\grave{e})\, Q_i(\grave{e}) \qquad [6\text{-}3]$$

where: a_i is the discrimination parameter for item *i*:

$P_i(\grave{e}) = 1/(1+ \mathrm{EXP}(-a_i(\grave{e} - b_i)))$,

$Q_i(\grave{e}) = 1 - P_i(\grave{e})$,

$\grave{e}$ is the ability level of interest.

To illustrate the use of equation 6-3, the amount of item information will be computed at seven ability levels for an item having parameter values of $b = 1.0$ and $a = 1.5$.

$\grave{e}$	L	$EXP(-L)$	$P_i(\grave{e})$	$Q_i(\grave{e})$	$P_i(\grave{e})\,Q_i(\grave{e})$	a^2	$I_i(\grave{e})$
-3	-6	403.43	.00	1.00	.00	2.25	.00
-2	-4.5	90.02	.01	.99	.01	2.25	.02
-1	-3.0	20.09	.05	.95	.05	2.25	.11
0	-1.5	4.48	.18	.82	.15	2.25	.34
1	0.0	1.00	.50	.50	.25	2.25	.56
2	1.5	.22	.82	.18	.15	2.25	.34
3	3.0	.05	.95	.05	.05	2.25	.11

Table 6-1. Calculation of item information under a two-parameter model, $b = 1.0$, $a = 1.5$

This item information function increases rather smoothly as ability increases and reaches a maximum value of .56 at an ability of 1.0. After this point, it decreases. The obtained item information function is symmetrical about the value of the item's difficulty parameter. Such symmetry holds for all item information functions under one- and two-parameter models. When only a single item is involved and the discrimination parameter has a moderate value, the magnitude of the amount of item information is quite small.

One-Parameter Item Characteristic Curve Model

Under a one-parameter (Rasch) model, the item information is defined as:

$$I_i(\grave{e}) = P_i(\grave{e})\, Q_i(\grave{e}) \qquad [6\text{-}4]$$

This is exactly the same as that under a two-parameter model when the value of the discrimination parameter is set to 1. To illustrate the use of equation 6-4, the amount of item information will be calculated for an item having a difficulty parameter of 1.0.

$\grave{e}$	L	$EXP(-L)$	$P_i(\grave{e})$	$Q_i(\grave{e})$	$P_i(\grave{e})\, Q_i(\grave{e})$	a^2	$I_i(\grave{e})$
-3	-4.0	45.60	.02	.98	.02	1	.02
-2	-3.0	20.09	.05	.95	.05	1	.05
-1	-2.0	7.39	.12	.88	.11	1	.11
0	-1.0	2.72	.27	.73	.20	1	.20
1	0.0	1.00	.50	.50	.25	1	.25
2	1.0	.37	.73	.27	.20	1	.20
3	2.0	.14	.88	.12	.11	1	.11

Table 6-2. Calculation of the item information under the Rasch model, $b = 1.0$

The general level of the amount of information yielded by this item is somewhat lower than that of the previous example. This is a reflection of the value of the item discrimination parameter being smaller than that of the previous item. Again, the item information function is symmetric about the value of the difficulty parameter.

Three-Parameter Item Characteristic Curve Model

In Chapter 2, it was mentioned that the three-parameter model does not possess the nice mathematical properties of the logistic function. The loss of these properties becomes apparent in the complexity of the equation given below for the amount of item information under this model.

$$I_i(q) = a^2\left[\frac{Q_i(q)}{P_i(q)}\right]\left[\frac{P_i(q) - c^2}{(1 - c^2)}\right] \qquad [6\text{-}5]$$

where: $P_i(\grave{e}) = c + (1 - c)\,(1/(1 + EXP(-L)))$ and $L = a(\grave{e} - b)$
$Q_i = 1.0 - P_i(\grave{e})$.

To illustrate the use of these formulas, the computations will be shown for an item having parameter values of $b = 1.0$, $a = 1.5$, $c = .2$. The values of b and a are the same as those for the preceding two-parameter example. The computations will be performed in detail at an ability level of $\grave{e} = 0.0$.

$L = 1.5\,(0 - 1) = -1.5$

$EXP(-L) = 4.482$

$1/(1 + EXP(-L)) = .182$

$P_i(\grave{e}) = c + (1 - c)\,(1/(1 + EXP(-L))) = .2 + .8\,(.182) = .346$

$Q_i(\grave{e}) = 1 - .346 = .654$

$Q_i(\grave{e})/P_1(\grave{e}) = .654/.346 = 1.890$

$$(P_i(\grave{e}) - c)^2 = (.346 - .2)^2 = (.146)^2 = .021$$

$$(1 - c)^2 = (1 - .2)^2 = (.8)^2 = .64$$

$$a^2 = (1.5)^2 = 2.25$$

Then:

$$I_i(\grave{e}) = (2.25)(1.890)(.021)/(.64) = .142$$

Clearly, this is more complicated than the computations for the previous two models, which are, in fact, logistic models. The amount of item information computations for this item at seven ability levels is shown below.

$\grave{e}$	L	$P_i(\grave{e})$	$Q_i(\grave{e})$	$P_i(\grave{e})\,Q_i(\grave{e})$	$(P_i(\grave{e}) - c)$	$I_i(\grave{e})$
-3	-6.0	.20	.80	3.950	.000	.000
-2	-4.5	.21	.79	3.785	.000	.001
-1	-3.0	.24	.76	3.202	.001	.016
0	-1.5	.35	.65	1.890	.021	.142
1	0.0	.60	.40	.667	.160	.375
2	1.5	.85	.15	.171	.428	.257
3	3.0	.96	.04	.040	.481	.082

Table 6-3. Calculations for the amount of item information under a three-parameter model, $b = 1.0$, $a = 1.5$, $c = .2$

The shape of this information function is very similar to that for the preceding two-parameter example in which $b = 1.0$ and $a = 1.5$. However, the general level of the values for the amount of information is lower. For example, at an ability level of $\grave{e} = 0$, the item information was .142 under a three-parameter model and .34 under a two-parameter model having the same values of b and a. In addition, the maximum of the information function did not occur at an ability level corresponding to

the value of the difficulty parameter. The maximum occurred at an ability level slightly higher than the value of *b*. Because of the presence of the terms (1 - c) and ($P_i(\hat{e}\ )$ - c) in equation 6-5, the amount of information under a three-parameter model will be less than under a two-parameter model having the same values of *b* and *a*. When they share common values of *a* and *b*, the information functions will be the same when $c = 0$. When $c > 0$, the three-parameter model will always yield less information. Thus, the item information function under a two-parameter model defines the upper bound for the amount of information under a three-parameter model. This is reasonable, because getting the item correct by guessing should not enhance the precision with which an ability level is estimated.

Computing a Test Information Function

Equation 6-2 defined the test information as the sum of the amount of item informations at a given ability level. Now that the procedures for calculating the amount of item information have been shown for the three item characteristic curve models, the test information function for a test can be computed. To illustrate this process, a five-item test will be used. The item parameters under a two-parameter model are as follows:

Item	b	a
1	-1.0	2.0
2	-0.5	1.5
3	-0.0	1.5
4	0.5	1.5
5	1.0	2.0

The amount of item information and the test information will be computed for the same seven ability levels used in the previous examples.

	Item Information					
						Test
è	**1**	**2**	**3**	**4**	**5**	**Information**
-	.071	.051	.024	.012	.001	.159

-	.420	.194	.102	.051	.010	.777
-	1.000	.490	.336	.194	.071	2.091
0	.420	.490	.563	.490	.420	2.383
1	.071	.194	.336	.490	1.000	2.091
2	.010	.051	.102	.194	.420	.777
3	.001	.012	.024	.051	.071	.159

Table 6-4. Calculations for a test information function based upon five items

Each of the item information functions was symmetric about the value of the item's difficulty parameter. The five item discriminations had a symmetrical distribution around a value of 1.5. The five item difficulties had a symmetrical distribution about an ability level of zero. Because of this, the test information function also was symmetric about an ability of zero. The graph of this test information function is shown in Figure 6-4.

The graph of the test information function shows that the amount of information was relatively flat over the range $\grave{e} = -1$ to $\grave{e} = +1$; outside of this range, the amount of information decreased rather rapidly. However, in Table 6-4, the values of the test information varied over the whole ability scale. The apparent flat section of the plotted test information function is due to the coarseness of the information scale in the graph.

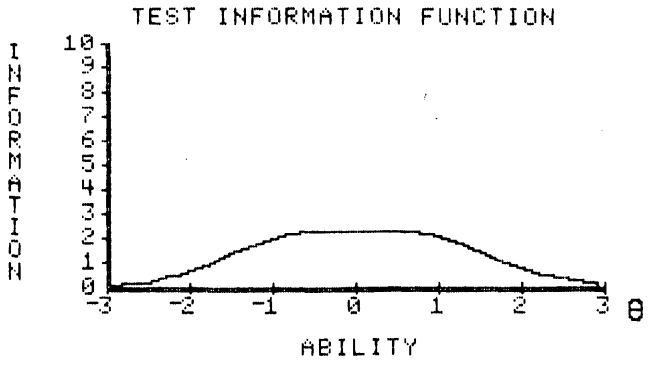


FIGURE 6-4. Test information function for the five items of Table 6-4

Interpreting the Test Information Function

While the shape of the desired test information function depends upon the purpose for which a test is designed, some general interpretations can be made. A test information function that is peaked at some point on the ability scale measures ability with unequal precision along the ability scale. Such a test would be best for estimating the ability of examinees whose abilities fall near the peak of the test information function. In some tests, the test information function is rather flat over some region of the ability scale. Such tests estimate some range of ability scores with nearly equal precision and outside this range with less precision. Thus, the test would be a desirable one for those examinees whose ability falls in the given range. When interpreting a test information function, it is important to keep in mind the reciprocal relationship between the amount of information and the variability of the ability estimates. To translate the amount of information into a standard error of estimation, one need only take the reciprocal of the square root of the amount of test information.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad [6\text{-}6]$$

For example, in Figure 6-4, the maximum amount of test information was 2.383 at an ability level of 0.0. This translates into a standard error of .65, which means roughly that 68 percent of the estimates of this ability level fall between -.65 and +.65. Thus, this ability level is estimated with a modest amount of precision.

Computer Session for Chapter 6

The purpose of this computer session is to enable you to develop a sense of how the form of the test information function depends upon the parameters of the items constituting the test. You will establish the parameter values for the items in a small test, then the computer will display the test information function on the screen. You can try different item characteristic curve models to determine how the choice of model affects the shape of the test information function. Under each model, different mixes of item parameter values can be used and the resultant test information function obtained. You should reach the point where you can predict the form of the test information function from the values of the item parameters.

Procedures for an Example Case

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight the TEST CHARACTERISTIC CURVE session and click on [CONTINUE]. The TEST SPECIFICATION screen will appear.

c. Click on [NUMBER OF ITEMS] and set the number of items in the test to N = 10.

d. In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on TWO PARAMETER.

e. In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

f. Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

g. Click on [ENTER PARAMETERS] and then set the following item parameter values:

b = -.4,	a = 1.0
b = -.3,	a = 1.5
b = -.2,	a = 1.2
b = -.1,	a = 1.3
b = 0,	a = 1.0
b = 0,	a = 1.6
b = .1,	a = 1.6
b = .2,	a = 1.4
b = .3,	a = 1.1
b = .4,	a = 1.7

a. When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

b. Click on [CONTINUE]. The test characteristic curve shown below will appear on the screen.

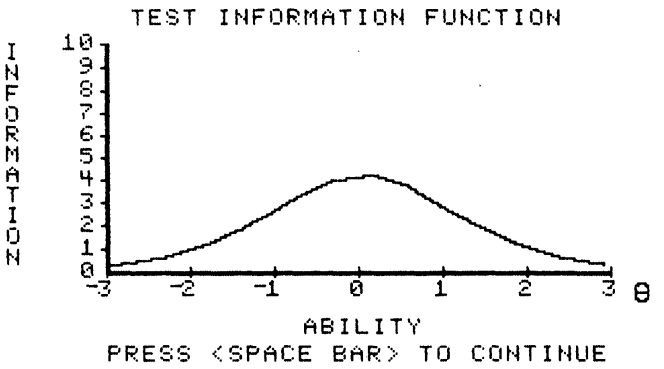


FIGURE 6-5. Test information function

j. The test information function is symmetric about an ability level of 0.0, reflecting the distribution of the item difficulties around zero. The maximum value of the amount of test information is approximately 4.2, which yields a standard error of estimate of .49. Within the range of ability from -1.0 to +1.0, the amount of test information is greater than 2.5, and the standard error of estimate is less than .63 in this range. Outside of this range, the amount of information is smaller, and at an ability level of -2.0 or +2.0, it is only about 1.0. At these points, the standard error of estimate is 1.0. Since this test has only ten items, the general level of the test information function is at a modest value, and the precision reflects this.

k. Click on [CONTINUE]. The SELECT OPTION FROM LIST screen will appear.

l. If you click on MODIFY EXISTING TEST, the ITEM PARAMETERS screen will appear and you can edit any of the values. If you click on CREATE NEW TEST, the TEST SPECIFICATION screen will appear. Click on CREATE NEW TEST.

Exercises

a. Using a two-parameter model

Exercise 1

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on TWO PARAMETER.

(4) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and then set all the item difficulty parameters to $b = 0.0$ and use various values of a that are all greater than 1.0 but less than 1.7.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE] and the test information function curve will appear on the screen.

(9) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.

(10) Click on CREATE NEW TEST.

Exercise 2

(11) The TEST SPECIFICATION screen will appear.

(12) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

(13) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on TWO PARAMETER.

(14) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(15) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(16) Click on [ENTER PARAMETERS] and then set all the item difficulty parameters to $b = 0.0$ and use various values of a that are all less than 1.0.

(17) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(18) Click on [CONTINUE] and the test information function curve will appear on the screen.

(19) The test information function will be symmetric about zero, but will have a much lower overall level than the previous test information function.

(20) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.

(21) Click on CREATE NEW TEST.

Exercise 3

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on TWO PARAMETER.

(4) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and then set all the item difficulty parameters to $b = 0.0$ and use various values of a that are all greater than 1.7. The maximum value you can use is 2.0.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE] and the test information function curve will appear on the screen.

(9) The test information function will have a maximum greater than that of all of the previous examples, thus illustrating the dependence of the amount of information upon the values of the discrimination parameter.

(10) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.
(11) Click on CREATE NEW TEST.

Exercise 4

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 5$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on TWO PARAMETER.

(4) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and then set the item parameters to values of your choice.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE] and the test information function curve will appear on the screen.

(9) The general level of the test information function will be much lower than the corresponding example. Depending on how you chose the values of *b* and *a*, the shape of the curve could be quite similar to the previous case.

(10) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.

(11) Click on CREATE NEW TEST.

b. Using a Rasch model

Exercise 1

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on RASCH.

(4) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and then set all the item difficulty parameters to some common value other than zero.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE] and the test information function curve will appear on the screen.

(9) The test information curve will be centered on this common value. The general level of the amount of information will be modest because the Rasch model fixes the discrimination parameter at 1.0.

(10) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.

(11) Click on CREATE NEW TEST.

Exercise 2

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on RASCH.

(4) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and then set the item difficulty parameters to some values that are equally spaced over the full range of ability from -3 to +3.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE] and the test information function curve will appear on the screen.

(9) The test information function will be rather flat, and the general amount of information will be rather low.

(10) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.

(11) Click on CREATE NEW TEST.

c. Using a three-parameter model

Exercise 1

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on THREE PARAMETER.

(4) In the SELECT ITEM PARAMETER CREATION METHOD list, click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and then select values of b and a that vary in value. Set the value of $c = .1$ for all items. Write down the values of b and a so they can be used again.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE] and the test information function curve will appear on the screen.

(9) Take note of the shape and general level of the obtained test information function.

(10) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.

(11) Click on CREATE NEW TEST.

Exercise 2

(1) The TEST SPECIFICATION screen will appear.

(2) Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

(3) In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on THREE PARAMETER.

(4) In the SELECT ITEM PARAMETER CREATION METHOD, list click on USER INPUT OF ITEM PARAMETER VALUES.

(5) Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

(6) Click on [ENTER PARAMETERS] and then use the same values of b and a that were used in the previous problem, but set all the values of $c = .35$.

(7) When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

(8) Click on [CONTINUE] and the test information function curve will appear on the screen.

(9) The resulting test information function will have a shape similar to that of the previous problem. However, the general level of the amount of test information will be less than that of the previous example. This illustrates the effect of guessing upon the precision with which ability is estimated.

(10) Click on [CONTINUE] and the SELECT OPTION FROM LIST screen will appear.

(11) Click on CREATE NEW TEST.

d. Exploratory exercises

1. Use a model of your choice and select values of the item parameters such that the test information function approximates a horizontal line. Use a ten-item test.

2. Experiment a bit with different item characteristic curve models, parameter values, and number of items. To make things easier, let the computer select the values of the item parameters by responding to the message SELECT ITEM

PARAMETER CREATION METHOD by clicking on COMPUTER GENERATED ITEM PARAMETER VALUES.

It will be helpful to make rough sketches of the test information functions displayed and notes to indicate the nature of the mix of item parameter values. The goal is to be able to predict what the form of the test information function will be from the values of the item parameters.

Things To Notice

1. The general level of the test information function depends upon:

    a. The number of items in the test.
    b. The average value of the discrimination parameters of the test items.
    c. Both of the above hold for all three item characteristic curve models.

2. The shape of the test information function depends upon:

    a. The distribution of the item difficulties over the ability scale.
    b. The distribution and the average value of the discrimination parameters of the test items.

3. When the item difficulties are clustered closely around a given value, the test information function is peaked at that point on the ability scale. The maximum amount of information depends upon the values of the discrimination parameters.

4. When the item difficulties are widely distributed over the ability scale, the test information function tends to be flatter than when the difficulties are tightly clustered.

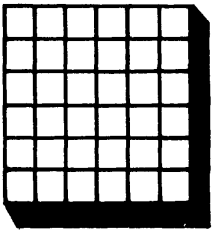
5. Values of $a < 1.0$ result in a low general level of the amount of test information.

6. Values of $a > 1.7$ result in a high general level of the amount of test information.

7. Under a three-parameter model, values of the guessing parameter c greater than zero lower the amount of test information at the low-ability levels. In addition, large values of c reduce the general level of the amount of test information.

8. It is difficult to approximate a horizontal test information function. To do so, the values of b must be spread widely over the ability scale,

and the values of *a* must be in the moderate to low range and have a U-shaped distribution.



CHAPTER 7
Test Calibration

CHAPTER 7

Test Calibration

For didactic purposes, all of the preceding chapters have assumed that the metric of the ability scale was known. This metric had a midpoint of zero, a unit of measurement of 1, and a range from negative infinity to positive infinity. The numerical values of the item parameters and the examinee's ability parameters have been expressed in this metric. While this has served to introduce you to the fundamental concepts of item response theory, it does not represent the actual testing situation. When test constructors write an item, they know what trait they want the item to measure and whether the item is designed to function among low-, medium- or high-ability examinees. But it is not possible to determine the values of the item's parameters *a priori*. In addition, when a test is administered to a group of examinees, it is not known in advance how much of the latent trait each of the examinees possesses. As a result, a major task is to determine the values of the item parameters and examinee abilities in a metric for the underlying latent trait. In item response theory, this task is called test calibration, and it provides a frame of reference for interpreting test results. Test calibration is accomplished by administering a test to a group of M examinees and dichotomously scoring the examinees' responses to the N items. Then mathematical procedures are applied to the item response data in order to create an ability scale that is unique to the particular combination of test items and examinees. Then the values of the item parameter estimates and the examinees' estimated abilities are expressed in this metric. Once this is accomplished, the test has been calibrated, and the test results can be interpreted via the constructs of item response theory.

The Test Calibration Process

The technique used to calibrate a test was proposed by Alan Birnbaum in 1968 and has been implemented in widely used computer programs such as BICAL (Wright and Mead, 1976) and LOGIST (Wingersky, Barton and Lord, 1982). The Birnbaum paradigm is an iterative procedure employing two stages of maximum likelihood estimation. In one stage, the parameters of the N items in the test are estimated, and in the second

stage, the ability parameters of the *M* examinees are estimated. The two stages are performed iteratively until a stable set of parameter estimates is obtained. At this point, the test has been calibrated and an ability scale metric defined.

Within the first stage of the Birnbaum paradigm, the estimated ability of each examinee is treated as if it is expressed in the true metric of the latent trait. Then the parameters of each item in the test are estimated via the maximum likelihood procedure discussed in Chapter 3. This is done one item at a time, because an underlying assumption is that the items are independent of each other. The result is a set of values for the estimates of the parameters of the items in the test.

The second stage assumes that the item parameter estimates yielded by the first stage are actually the values of the item parameters. Then, the ability of each examinee is estimated using the maximum likelihood procedure presented in Chapter 5. It is assumed that the ability of each examinee is independent of all other examinees. Hence, the ability estimates are obtained one examinee at a time.

The two-stage process is repeated until some suitable convergence criterion is met. The overall effect is that the parameters of the *N* test items and the ability levels of the *M* examinees have been estimated simultaneously, even though they were done one at a time. This clever paradigm reduces a very complex estimation problem to one that can be implemented on a computer.

The Metric Problem

An unfortunate feature of the Birnbaum paradigm is that it does not yield a unique metric for the ability scale. That is, the midpoint and the unit of measurement of the obtained ability scale are indeterminate; i.e., many different values work equally well. In technical terms, the metric is unique up to a linear transformation. As a result, it is necessary to "anchor" the metric via arbitrary rules for determining the midpoint and unit of measurement of the ability scale. How this is done is up to the persons implementing the Birnbaum paradigm in a computer program. In the BICAL computer program, this anchoring process is performed after the first stage is completed. Thus, each of two stages within an

iteration is performed using a slightly different ability scale metric. As the overall iterative process converges, the metric of the ability scale also converges to a particular midpoint and unit of measurement. The crucial feature of this process is that the resulting ability scale metric depends upon the specific set of items constituting the test and the responses of a particular group of examinees to that test. It is not possible to obtain estimates of the examinee's ability and of the item's parameters in the true metric of the underlying latent trait. The best we can do is obtain a metric that depends upon a particular combination of examinees and test items.

Test Calibration Under the Rasch Model

There are three different item characteristic curve models to choose from and several different ways to implement the Birnbaum paradigm. From these, the author has chosen to present the approach based upon the one-parameter logistic (Rasch) model as implemented by Benjamin Wright and his co-workers in the BICAL computer program. Under this model, each item has only one parameter to be estimated. The procedures work well with small numbers of test items and small numbers of examinees. The metric anchoring procedure is simple, and the basic ideas of test calibration are easy to present.

The calibration of a ten-item test administered to a group of 16 examinees will be used below to illustrate the process. The information presented is based upon the analysis of Data Set 1 contained in the computer session CALIBRATE A TEST on the companion Web site. You may elect to work through this section in parallel with the computer session, but it is not necessary because all the computer displays will be presented in the text.

The ten-item test is one that has been matched to the average ability of a group of 16 examinees. The examinees' item responses have been dichotomously scored, 1 for correct and 0 for incorrect. The goal is to use this item response data to calibrate the test. The actual item response vectors for each examinee are presented below, and each row represents the item responses made by a given examinee.

ITEM RESPONSES BY EXAMINEE

MATCHED TEST ITEM

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**	**RS**
	01	1		1					1			2
	02	1		1								2
	03	1	1	1		1		1				5
	04	1	1	1		1						4
E	05					1						1
X	06	1	1		1							3
A	07	1					1	1	1			4
M	08	1				1	1			1		4
I	09	1		1			1			1		4
N	10	1				1					1	3
E	11	1	1		1	1	1	1	1	1	1	9
E	12	1	1	1	1	1	1	1	1	1		9
	13	1	1	1		1		1			1	6
	14	1	1	1	1	1	1	1	1	1		9
	15	1	1		1	1	1	1	1	1	1	9
	16	1	1	1	1	1	1	1	1	1	1	10

Table 7-1. Item responses by examinee

In Chapter 5 it was observed that it is impossible to estimate an examinee's ability if he or she gets none or all of the test items correct. Inspection of Table 7-1 reveals that examinee 16 answered all of the items correctly and must be removed from the data set. Similarly, if an item is answered correctly by all of the examinees or by none of the examinees, its item difficulty parameter cannot be estimated. Hence, such an item must be removed from the data set. In this particular example,

no items were removed for this reason. One of the unique features of test calibration under the Rasch model is that all examinees having the same number of items correct (the same raw score) will obtain the same estimated ability. As a result, it is not necessary to distinguish among the several examinees having the same raw test score. Consequently, rather than use the individual item responses, all that is needed is the number of examinees at each raw score answering each item correctly. Because of this and the removing of items, an edited data set is used as the initial starting point for test calibration procedures under the Rasch model. The edited data set for this example is presented below.

FREQUENCY COUNTS FOR EDITED DATA
ELIMINATED EXAMINEES #16
ELIMINATED ITEMS # NONE

		ITEM									**1**	**Row**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
	1					1						1
S	2	1		2					1			4
C	3	2	1		1	1				1		6
O	4	4	1	2		2	3	1	1	2		16
R	5	1	1	1		1		1				5
E	6	1	1	1		1		1			1	6
	9	4	4	2	4	4	4	4	4	4	2	36
COL		13		8		10		7		7		
Total			8		5		7		6		3	74

Table 7-2. Frequency counts for the edited data

In Table 7-2, the rows are labeled by raw test scores ranging from 1 to 9. The row marginals are the total number of correct responses made by examinees with that raw test score. The columns are labeled by the item number from 1 to 10. The column marginals are the total number of

correct responses made to the particular item by the remaining examinees. (The double row of column totals was necessary to work around space limitations of the monitor screen.) Under the Rasch model, the only information used in the Birnbaum paradigm are the frequency totals contained in the row and column marginals. This is unique to this model and results in simple computations within the maximum likelihood estimation procedures employed at each stage of the overall process.

Given the two frequency vectors, the estimation process can be implemented. Initial estimates are obtained for the item difficulty parameters in the first stage, and the metric of the ability scale must be anchored. Under the Rasch model, the anchoring procedure takes advantage of the fact that the item discrimination parameter is fixed at a value of 1 for all items in the test. Because of this, the unit of measurement of the estimated abilities is fixed at a value of 1. All that remains, then, is to define the midpoint of the scale. In the BICAL computer program, the midpoint is defined as the mean of the estimated item difficulties. In order to have a convenient midpoint value, the mean item difficulty is subtracted from the value of each item's difficulty estimate, resulting in the rescaled mean item difficulty having a value of zero. Because the item difficulties are expressed in the same metric as the ability scale, the midpoint and unit of measurement of the latter have now been determined. Since this is done between stages, the abilities estimated in the second stage will be in the metric defined by the rescaled item parameter estimates obtained in the first stage. The ability estimate corresponding to each raw test score is obtained in the second stage using the rescaled item difficulties as if they were the difficulty parameters and the vector of row marginal totals. The output of this stage is an ability estimate for each raw test score in the data set. At this point, the convergence of the overall iterative process is checked. In the BICAL program, Wright summed the absolute differences between the values of the item difficulty parameter estimates for two successive iterations of the paradigm. If this sum was less than .01, the estimation process was terminated. If it was greater than .01, then another iteration was performed and the two stages were done again. Thus, the process of stage one, anchor the metric, stage two, and check for convergence is repeated until the criterion is met. When this happens, the current values of the item and ability parameter estimates are accepted and an ability

scale metric has been defined. The estimates of the item difficulty parameters for the present example are presented below.

DATA SET 1
ITEM PARAMETER ESTIMATES

Item	Difficulty
1	-2.37
2	-0.27
3	-0.27
4	+0.98
5	-1.00
6	+0.11
7	+0.11
8	+0.52
9	+0.11
10	+2.06

Table 7-3. Estimated item difficulty parameters

You can verify that the sum of the item difficulties is zero (within rounding errors). The interpretation of the values of the item parameter estimates is exactly that presented in Chapter 2. For example, item 1 has an item difficulty of -2.37, which locates it at the low end of the ability scale. Item 6 has a difficulty of +.11, which locates it near the middle of the ability scale. Item 10 has a difficulty of 2.06, which locates it at the high end of the ability scale. Thus, the usual interpretation of item difficulty as locating the item on the ability scale holds. Because of the anchoring procedures used, these values are actually relative to the average item difficulty of the test for these examinees.

Although an ability estimate has been reported in Table 7-4 for each examinee, all examinees with the same raw score obtained the same ability estimate. For example, examinees 1 and 2 both had raw scores of 2

and obtained an estimated ability of -1.5. Examinees 7, 8 and 9 had raw scores of 4 and shared a common estimated ability of -.42. This unique feature is a direct consequence of the fact that, under the Rasch model, the value of the discrimination parameter is fixed at 1 for all of the items in the test. This aspect of the Rasch model is appealing to practitioners because they intuitively feel that examinees obtaining the same raw test score should receive the same ability estimate. When the two- and three-parameter item characteristic curve models are used, an examinee's ability estimate depends upon the particular pattern of item responses rather than the raw score. Under these models, examinees with the same item response pattern will obtain the same ability estimate. Thus, examinees with the same raw score could obtain different ability estimates if they answered different items correctly.

DATA SET 1
ABILITY ESTIMATION

Examinee	Obtained	Raw Score
1	-1.50	2
2	-1.50	2
3	+0.02	5
4	-0.42	4
5	-2.37	1
6	-0.91	3
7	-0.42	4
8	-0.42	4
9	-0.42	4
10	-0.91	3
11	+2.33	9
12	+2.33	9
13	+0.46	6
14	+2.33	9

15	+2.33	9
16	*****	10

Table 7-4. Obtained ability estimates

Examinee number 16 was not included in the computations due to being removed because of a perfect raw score. The ability estimate obtained by a given examinee is interpreted in terms of where it locates the examinee on the ability scale. For example, examinee number 7 had an estimated ability of -.42, which places him or her just below the midpoint of the scale. The ability estimates can be treated just like any other score. Their distribution over the ability scale can be plotted, and the summary statistics of this distribution can be computed. In the present case, this yields a mean of .06 and a standard deviation of 1.57. Thus, examinee number 7 had an ability score that was about .27 standard deviations below the mean ability of the group. However, one would not typically interpret an examinee's ability score in terms of the distribution of the scores for the group of examinees. To do so is to ignore the fact that the ability score can be interpreted directly as the examinee's position on the ability scale.

Summary of the Test Calibration Process

The end product of the test calibration process is the definition of an ability scale metric. Under the Rasch model, this scale has a unit of measurement of 1 and a midpoint of zero. Superficially this looks exactly the same as the ability scale metric used in previous chapters. However, it is not the metric of the underlying latent trait. The obtained metric depends upon the item responses yielded by a particular combination of examinees and test items being subjected to the Birnbaum paradigm. Since the true metric of the underlying latent trait cannot be determined, the metric yielded by the Birnbaum paradigm is used as if it were the true metric. The obtained item difficulty values and the examinee's ability are interpreted in this metric. Thus, the test has been calibrated. The outcome of the test calibration procedure is to locate each examinee and item along the obtained ability scale. In the present example, item 5 had a difficulty of -1 and examinee 10 had an ability estimate of -.91. Therefore, the probability of examinee 10 answering item 5 correctly is

approximately .5. The capability to locate items and examinees along a common scale is a powerful feature of item response theory. This feature allows one to interpret the results of a test calibration within a single framework and provides meaning to the values of the parameter estimates.

Computer Session for Chapter 7

This computer session is a bit different from those of the previous chapters. Because it would be difficult for you to create data sets to be calibrated, three sets have been prestored on the Web site. Each of these will be used to calibrate a test, and the results will be displayed on the screen. You will simply step through each of the data sets and calibration results. There are some definite goals in this process. First, you will become familiar with the input data and how it is edited. Second, the item difficulty estimates and the examinee's ability estimates can be interpreted. Third, the test characteristic curve and test information functions for the test will be shown and interpreted.

Three different ten-item tests measuring the same latent trait will be used. A common group of 16 examinees will take all three of the tests. The tests were created so that the average difficulty of the first test was matched to the mean ability of the common group of examinees. The second test was created to be an easy test for this group. The third test was created to be a hard test for this group. Each of these test-group combinations will be subjected to the Birnbaum paradigm and calibrated separately. There are two reasons for this approach. First, it illustrates that each test calibration yields a unique metric for the ability scale. Second, the results can be used to show the process by which the three sets of test results can be placed on a common ability scale.

Procedures for the test calibration session

a. Data set 1

This ten-item test has a mean difficulty that is matched to the average ability of the group of 16 examinees.

(1) Follow the start-up procedures described in the Introduction.

(2) Use the mouse to highlight the CALIBRATE A TEST session and click on [SELECT].

(3) Read the explanatory screens and click on [CONTINUE] to move from one screen to the next.

(4) The table of item response vectors will be displayed. This will be the same as Table 7-1. Notice that examinee 16 answered all items correctly. Click on [CONTINUE].

(5) The table of edited data will be displayed. It will be the same as Table 7-2. Notice that examinee 16 has been eliminated and that no items were eliminated. Click on [CONTINUE].

(6) A screen indicating that the Birnbaum paradigm has been used to calibrate the test will be shown. Click on [CONTINUE].

(7) The table of item difficulty estimates for test 1 will be shown. This is the same as Table 7-3. Click on [CONTINUE].

(8) The estimated abilities of the 16 examinees and their raw scores will be shown. The screen will be the same as Table 7-4. The ability estimates had a mean of .062 and a standard deviation of 1.57. Notice that examinee 16 did not receive an ability estimate.

(9) The message DO YOU WANT TO REVIEW DATA SET 1 RESULTS AGAIN? appears. If you click on the YES button, you will be returned to step 4. If you click on the NO button, the next screen will appear. Click on the NO button.

(10) A NO response will result in the test characteristic curve being displayed. Take note of the fact that the mid-true score (a true score equal to one-half the number of items) corresponds to an ability level of zero. This reflects the anchoring procedure that sets the average item difficulty to zero. Click on [CONTINUE].

(11) The test information function will be displayed next. The curve is reasonably symmetric and has a well-defined hump in the middle. The form of the curve indicates that ability is estimated with the greatest precision in the neighborhood of the middle of the ability scale. The peak of the test information function occurs at a point slightly above the midpoint of the ability scale. This reflects the distribution of the item difficulties, as there were six items with positive values and only four with negative values. Thus, there is a very slight emphasis upon positive ability levels.

(12) Clicking on [DISPLAY FIRST CURVE] will cause the graph of the test characteristic curve to reappear. This will allow you to alternate between the Test Characteristic Curve and Test Information Function screens.

(13) To continue the session, respond to the question, DO NEXT DATA SET? by clicking on the YES button.

b. Data set 2

This ten-item test was constructed to be an easy test for the common group of 16 examinees. Since the computer procedures for this data set will be exactly the same as for data set 1, they will not be repeated in detail. Only the significant results will be noted.

(1) In the display of the edited data, examinees 15 and 16 have been eliminated for having perfect raw scores.

(2) The mean of the estimated item difficulties is .098, which is close to zero. Six of the items obtained positive item difficulties, and the distribution of the difficulties is somewhat U-shaped.

(3) The ability estimates had a mean of .44 and a standard deviation of 1.35. It is interesting to note that examinee 9 had a raw score of 4 on the first test and obtained an estimated ability of -.42. On the second test, the raw score was 7 and the

ability estimate was 1.02. Yet the examinee's true ability is the same in both cases.

(4) The mid-true score of the test characteristic curve again corresponds to an ability level of zero. The form of the test characteristic curve is nearly identical to that of the first test.

(5) The test information function is symmetric and has a somewhat rounded appearance. The maximum amount of information occurred at an ability level of roughly .5.

(6) Respond to the message DO NEXT DATA SET? by clicking on the YES button.

c. Data set 3

This ten-item test was constructed to be a hard test for the common group of 16 examinees. Because the computer procedures will be the same as for the previous two examples, only the results of interest will be discussed.

(1) Inspection of the table of item response vectors shows that examinees 1 and 3 have raw scores of zero and will be removed. Inspection of the columns reveals that none of the examinees answered item 10 correctly and it will be removed from the data set. In addition, after removing the two examinees, item 1 was answered correctly by all of the remaining examinees. Thus, this item must also be removed. Upon doing this, examinees 2 and 6 now have raw scores of zero because the only item they answered correctly was item 1. After removing these two additional examinees, no further editing is needed. Such multiple-stage editing is quite common in test calibrating. It should be noted that after editing, the data set is smaller than the previous two, and the range of raw scores is now from 1 to 7.

(2) The mean of the eight estimated item difficulties was .0013, which again is close to zero. Three of the items had positive values of difficulty estimates. Item 8 had a difficulty of 1.34,

while the remaining seven item difficulties fell in the range of -.67 to +.79.

(3) The 12 examinees used in the test calibration had a mean of -.11 and a standard deviation of 1.41.

(4) The test characteristic curve is similar to the previous two, and the mid-true score occurs again at an ability level of zero. But the upper part of the curve approaches a value of 8 rather than 10.

(5) The test information function was nearly symmetrical about an ability level of roughly zero. The curve was a bit less peaked than either of the two previous test information functions, and its maximum was slightly lower.

(6) Respond to the message DO NEXT DATA SET? by clicking on the NO button. This will result in termination of the session, and the main menu will reappear on the screen.

The reader should ponder a bit as to why the mean ability of the common group of examinees is not the same for all three calibrations. The item invariance principle says that they should all be the same. Is the principle wrong or is something else functioning here? The resolution of this inconsistency is presented after the Things To Notice section.

Things To Notice

1. In all three calibrations, examinees were removed in the editing process. As a result, the common group is not quite the same in each of the calibrations.

2. Although the tests were designed to represent tests that were easy, hard, and matched relative to the average ability of the common group, the results did not reflect this. Due to the anchoring process, all three test calibrations yielded a mean item difficulty of zero.

3. Within each calibration, examinees with the same raw test score obtained the same estimated ability. However, a given raw score will not yield the same estimated ability across the three calibrations.

4. Even though the same group of examinees was administered all three tests, the mean and standard deviations of their ability estimates were different for each calibration. This can be attributed to a number of causes. The primary reason is that due to the anchoring process, the value of the mean estimated abilities is expressed relative to the mean item difficulty of the test. Thus, the mean difficulty of the easy test should result in a positive mean ability. The mean ability on the hard test should have a negative value. The mean ability on the matched test should be near zero. The changing group membership also accounts for some of the differences, particularly when the group was small to start with. Finally, the overall amount of information is rather small in all three test information functions. Thus, the ability level of none of the examinees is being estimated very precisely. As a result, the ability estimate for a given examinee is not necessarily very close to his or her true ability.

5. The anchoring procedure set the mean item difficulty equal to zero, and thus the midpoint of the ability scale to zero. A direct consequence of this is that the mid-true score for all three test characteristic curves occurs at an ability level of zero. The similarity in the shapes of the curves for the first two data sets was due to the item difficulties being distributed in an approximately symmetrical manner around the zero point. The fact that all the items had the same value of the discrimination parameter (1.0) makes the slopes of

the first two curves similar. The curve for data set 3 falls below those for sets 1 and 2 because it was based on only eight items. However, its general shape is similar to the previous two curves, and its mid-true score occurred at an ability level of zero.

6. Although the test information functions were similar, there were some important differences. The curve for the matched test had the same general level as that for the easy test but was a bit flatter, indicating this test maintained its level of precision over a slightly wider range. The test information function for the hard test had a slightly smaller amount of information at its midpoint. Thus, it had a bit less precision at this point. However, the curve decreased a bit faster than the other two, indicating that the test did not hold its precision over a wide range of ability.

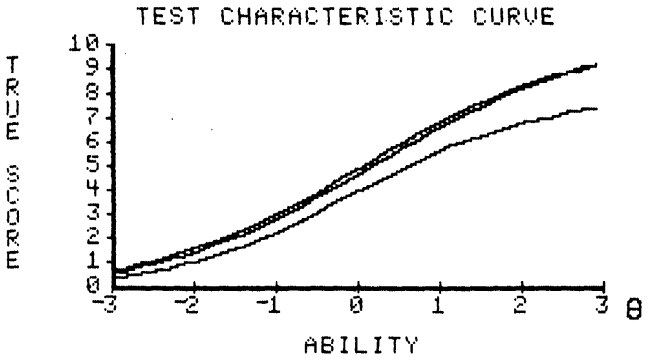


FIGURE 7-1. Test characteristic curves for the three data sets

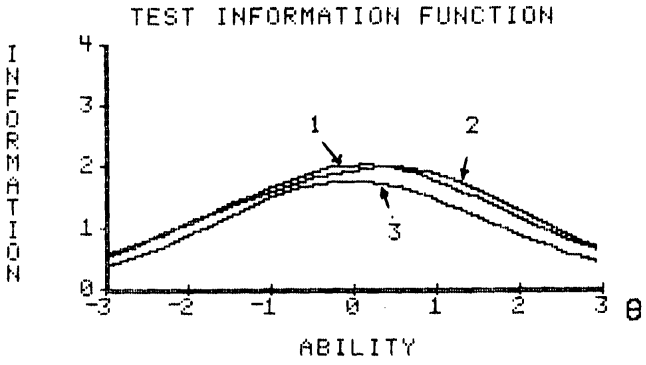


FIGURE 7-2. Test information function for the three data sets

Putting the Three Tests on a Common Ability Scale (Test Equating)

The principle of the item invariance of an examinee's ability indicates that an examinee should obtain the same ability estimate regardless of the set of items used. However, in the three test calibrations done above, this did not hold. The problem is not in the invariance principle, but in the test calibrations. The invariance principle assumes that the values of the item parameters of the several sets of items are all expressed in the same ability-scale metric. In the present situation, there are three different ability scales, one from each of the calibrations. Because of this, the same examinee will get three apparently different values of estimated ability rather than a common value. The intent of the three tests was to have one matched to the mean ability of the common group of 16 examinees, one to be easy for the group, and one to be hard for the group. Clearly, the average difficulties of these tests were intended to be different, but the anchoring process forced each test to have a mean item difficulty of zero. All is not lost, however, because forcing the mean item difficulty of the test to zero results in the average estimated ability of the group

reflecting the mean of the item difficulties before rescaling. Thus, what had originally been differences in average difficulty of the three tests now becomes differences in the mean ability of the common group of examinees. From the results presented above, the mean of the common group was .06 for the matched test, .44 for the easy test, and -.11 for the hard test. This tells us that the mean ability from the matched test is about what it should be. The mean from the easy test tells us that the average ability is above the mean item difficulty of the test, and this is as it should be. Finally, the mean ability from the hard test is below the mean item difficulty. Again, this is what one would expect. Since item difficulty and ability are measured in the same metric, we can use the mean abilities to position the tests on a common scale. The question then becomes "What scale?" and the choice becomes choosing which particular test calibration to use as the baseline. In the present case, the scale yielded by the calibration of the matched test and the common group is the most logical choice for a baseline metric. This calibration yielded a mean ability of .062 and a mean item difficulty of zero. In addition, we know one test was to be easy and one was to be hard. Thus, using the matched test calibration as the baseline seems appropriate. Because the Rasch model was used, the unit of measurement for all three calibrations is unity. Therefore, to bring the easy and hard test results to the baseline metric only involved adjusting for the differences in midpoints. In the paragraphs below, the results for the easy and hard tests will be transformed to the baseline metric.

Easy Test

The shift factor needed is the difference between the mean estimated ability of the common group on the easy test (.444) and on the matched test (.062), which is .382. To convert the values of the item difficulties for the easy test to baseline metric, one simply subtracts .382 from each item difficulty. The resulting values are shown in Table 7-5. Similarly, each examinee's ability can be expressed in the baseline metric by subtracting .382 from it. The transformed values are shown in Table 7-6 below.

Hard Test

The hard test results can be expressed in the baseline metric by using the differences in mean ability. The shift factor is -.111, -.062, or -.173. Again, subtracting this value from each of the item difficulty estimates puts them in the baseline metric. The transformed values are shown in Table 7-5. The ability estimates of the common group yielded by the hard test can be transformed to the baseline metric of the matched test. This was accomplished by using the same shift factor as was employed to rescale the item difficulty estimates. The results of rescaling each examinee's ability estimate to the baseline metric are reported in Table 7-6.

Item	Easy test	Matched test	Hard test
1	-1.492	-2.37	*****
2	-1.492	-.27	-.037
3	-2.122	-.27	-.497
4	-.182	.98	-.497
5	-.562	-1.00	.963
6	+.178	.11	-.497
7	.528	.11	.383
8	.582	.52	1.533
9	.880	.11	.443
10	.880	2.06	*****
	mean- .285	mean 0.00	mean .224

Table 7-5. Item difficulties in the baseline metric

After transformation, the mean item difficulties show the desired relations on the baseline ability scale. The matched test has a mean at the midpoint of the baseline ability scale. The easy test has a negative value, and the hard test has a positive value. The average difficulty of both tests

is about the same distance from the middle of the scale. In technical terms we have "equated" the tests, i.e., put them on a common scale.

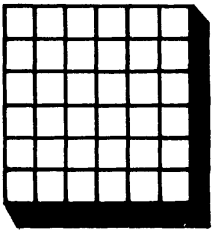
Item	Easy test	Matched test	Hard test
1	-2.900	-.1.50	*****
2	-.772	-1.50	*****
3	-1.962	.02	*****
4	-.292	-.42	-.877
5	-.292	-2.37	-.877
6	.168	-.91	*****
7	1.968	-.42	-1.637
8	.168	-.42	-.877
9	.638	-.42	-1.637
10	.638	-.91	-.877
11	.638	2.33	.153
12	1.188	2.33	.153
13	-.292	.46	.153
14	1.968	2.33	2.003
15	*****	2.33	1.213
16	*****	*****	2.003
mean	.062	.062	.062
Std. Dev.	1.344	1.566	1.413

Table 7-6. Ability estimates of the common group in the baseline metric

A number of interesting observations can be drawn from these results. The mean estimated ability of the common group was the same for all three tests. The standard deviations of the ability estimates were nearly the same for the easy and hard tests, and that for the matched test was

“in the ballpark.” Although the summary statistics were quite similar for all three sets of results, the ability estimates for a given examinee varied widely. The invariance principle has not gone awry; what you are seeing is sampling variation. The data set for each of the three test calibrations involved a small number of items (10) and a small number of examinees (16). As a result, the sampling variability of the item response data will be quite large, and one would not expect the several ability estimates to be the same. In Chapter 5, the reader was introduced to this concept. In this chapter, you are seeing it in a practical setting. Given the small size of the data sets, it is quite amazing that the results came out as nicely as they did. This demonstrates rather clearly the powerful capabilities of the Rasch model and Birnbaum’s maximum likelihood estimation paradigm as implemented in the BICAL computer program.

What was accomplished above is known in the field of psychometrics as test equating. All three of the tests have been placed on a common scale. After equating, the numerical values of the item parameters can be used to compare where different items function on the ability scale. The examinees’ estimated abilities also are expressed in this metric and can be compared. Although it has not been done here, it is also possible to compute the test characteristic curve and the test information function for the easy and hard tests in the baseline metric. Technically speaking, the tests were equated using the common group approach with tests of different difficulty. The ease with which test equating can be accomplished is one of the major advantages of item response theory over classical test theory.



CHAPTER 8
Specifying the Characteristics of a Test

CHAPTER 8

Specifying the Characteristics of a Test

During this transitional period in testing practices, many tests have been designed and constructed using classical test theory principles but have been analyzed via item response theory procedures. This lack of congruence between the construction and analysis procedures has kept the full power of item response theory from being exploited. In order to obtain the many advantages of item response theory, tests should be designed, constructed, analyzed, and interpreted within the framework of the theory. Consequently, the goal of this chapter is to provide the reader with experience in the technical aspects of test construction within the framework of item response theory.

Persons functioning in the role of test constructors do so in a wide variety of settings. They develop tests for commercial testing companies, governmental agencies, and school districts. In addition, teachers at all classroom levels develop tests to measure achievement. In all of these settings, the test construction process is usually based upon having a collection of items from which to select those to be included in a particular test. Such collections of items are known as item pools. Items are selected from such pools on the basis of both their content and their technical characteristics, i.e., their item parameter values. Under item response theory, a well-defined set of procedures is used to establish and maintain such item pools. A special name, item banking, has been given to these procedures. The basic goal is to have an item pool in which the values of the item parameters are expressed in a known ability-scale metric. If this is done, it is possible to select items from the item pool and determine the major technical characteristics of a test before it is administered to a group of examinees. If the test characteristics do not meet the design goals, selected items can be replaced by other items from the item pool until the desired characteristics are obtained. In this way, considerable time and money that would ordinarily be devoted to piloting the test are saved.

In order to build an item pool, it is necessary first to define the latent trait the items are to measure, write items to measure this trait, and pilot test the items to weed out poor items. After some time, a set of items measuring the latent trait of interest is available. This large set of items is then administered to a large group of examinees. An item characteristic curve model is selected, the examinees' item response data are analyzed via the Birnbaum paradigm, and the test is calibrated. The ability scale resulting from this calibration is considered to be the baseline metric of the item pool. From a test construction point of view, we now have a set of items whose item parameter values are known; in technical terms, a "precalibrated item pool" exists.

Developing a Test From a Precalibrated Item Pool

Since the items in the precalibrated item pool measure a specific latent trait, tests constructed from it will also measure this trait. While this may seem a bit odd, there are a number of reasons for wanting additional tests to measure the same trait. For example, alternate forms are routinely needed to maintain test security, and special versions of the test can be used to award scholarships. In such cases, items would be selected from the item pool on the basis of their content and their technical characteristics to meet the particular testing goals. The advantage of having a precalibrated item pool is that the parameter values of the items included in the test can be used to compute the test characteristic curve and the test information function before the test is administered. This is possible because neither of these curves depends upon the distribution of examinee ability scores over the ability scale. Thus, both curves can be obtained once the values of the item parameters are available. Given these two curves, the test constructor has a very good idea of how the test will perform before it is given to a group of examinees. In addition, when the test has been administered and calibrated, test equating procedures can be used to express the ability estimates of the new group of examinees in the metric of the item pool.

Some Typical Testing Goals

In order to make the computer exercises meaningful to you, several types of testing goals are defined below. These will then serve as the basis for specific types of tests you will create.

a. Screening tests.

   Tests used for screening purposes have the capability to distinguish rather sharply between examinees whose abilities are just below a given ability level and those who are at or above that level. Such tests are used to assign scholarships and to assign students to specific instructional programs such as remediation or advanced placement.

b. Broad-ranged tests.

   These tests are used to measure ability over a wide range of underlying ability scale. The primary purpose is to be able to make a statement about an examinee's ability and to make comparisons among examinees. Tests measuring reading or mathematics are typically broad-range tests.

c. Peaked tests.

   Such tests are designed to measure ability quite well in a region of the ability scale where most of the examinees' abilities will be located, and less well outside this region. When one deliberately creates a peaked test, it is to measure ability well in a range of ability that is wider than that of a screening test, but not as wide as that of a broad-range test.

Computer Session for Chapter 8

The purpose of this session is to assist you in developing the capability to select items from a precalibrated item pool to meet a specific testing goal. You will set the parameter values for the items of a small test in order to meet one of the three testing goals given above. Then the test

characteristic curve and the test information function will be shown on the screen and you can determine if the testing goal was met. If not, a new set of item parameters can be selected and the resultant curves obtained. With a bit of practice, you should become proficient at establishing tests having technical characteristics consistent with the design goals.

Some Ground Rules

a. It is assumed that the items would be selected on the basis of content as well as parameter values. For present purposes, the actual content of the items need not be shown.

b. No two items in the item pool possess exactly the same combination of item parameter values.

c. The item parameter values are subject to the following constraints:

$$-3 <= b <= +3$$

$$.50 <= a < +2.00$$

$$0 <= c <= .35$$

The values of the discrimination parameter have been restricted to reflect the range of values usually seen in well-maintained item pools.

Procedures for an Example Case

You are to construct a ten-item screening test that will separate examinees into two groups: those who need remedial instruction and those who don't, on the ability measured by the items in the item pool. Students whose ability falls below a value of -1 will receive the instruction.

a. Follow the start-up procedures described in the Introduction.

b. Use the mouse to highlight TEST SPECIFICATION, then click on [SELECT].

c. Read the explanatory screen and then click on [CONTINUE].

d. Click on [NUMBER OF ITEMS] and set the number of items in the test to $N = 10$.

e. In the SELECT ITEM CHARACTERISTIC CURVE MODEL list, click on TWO PARAMETER.

f. Respond to the question SETTINGS OK? by clicking on the YES button. The ITEM PARAMETERS screen will appear.

g. CLICK on [ENTER PARAMETERS] and then set the following item parameter values:

Item	Difficulty	Discrimination
1	b = -1.8	a = 1.2
2	b = -1.6	a = 1.4
3	b = -1.4	a = 1.1
4	b = -1.2	a = 1.3
5	b = -1.0	a = 1.5
6	b = - .8	a = 1.0
7	b = - .6	a = 1.4
8	b = - .4	a = 1.2
9	b = - .2	a = 1.1
10	b = 0.0	a = 1.3

The logic underlying these choices was one of centering the difficulties on the cutoff level of -1 and using moderate values of discrimination.

h. Study the table of item parameters for a moment. If you need to change a value, click on the value and the data input box will appear, allowing you to enter a new value.

i. When you are satisfied with the parameter values, respond to the message PARAMETER VALUES OK? by clicking on the YES button.

j. Click on [CONTINUE] and the test characteristic curve shown below will appear on the screen.

k. When the test characteristic curve appears on the screen, make note of the ability level at which the mid-true score occurs. Also note the slope of the curve at that ability level. The graph is shown here:

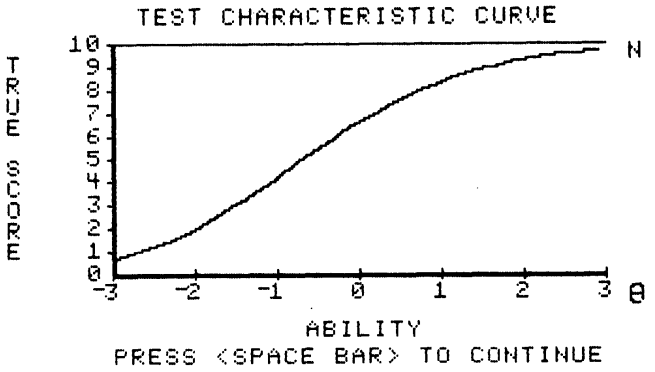


FIGURE 8-1. Test characteristic curve for the example

l. Click on [CONTINUE]. When the test information function appears on the screen, note the maximum amount of information and the ability level at which it occurred. The function is shown below.

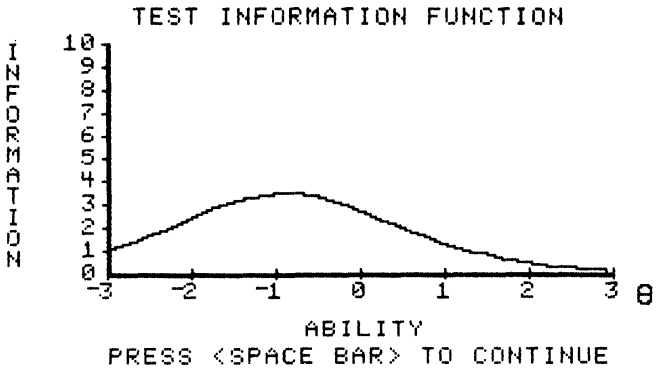


FIGURE 8-2. Test information function for the example

m. If you click on [DISPLAY FIRST CURVE], the test characteristic curve will appear again. Thus, you can alternate between the two graphs to study their relationship.

n. The design goal was to specify the items of a screening test that would function at an ability level of -1.0. In general, this goal has been met. The mid-true score corresponded to an ability level of -1.0. The test characteristic curve was not particularly steep at the cutoff level, indicating that the test lacked discrimination. The peak of the information function occurred at an ability level of -1.0, but the maximum was a bit small. The results suggest that the test was properly positioned on the ability scale but that a better set of items could be found.

The following changes would improve the test's characteristics: first, cluster the values of the item difficulties nearer the cutoff level; second, use larger values of the discrimination parameters. These two changes should steepen the test characteristic curve and increase the maximum amount of information at the ability level of -1.0.

o. When in the Test Information Function screen, click on [CONTINUE] and the next screen will appear.

p. Respond to the question DO ANOTHER TEST? by clicking on the YES button.

q. Respond to the question DO A NEW TEST? by clicking on the YES button.

r. Respond to the question PLOT ON SAME GRAPHS? by clicking on the YES button

s. Respond to the question SETTINGS OK? by clicking on the YES button.

t. Respond to the question PLOT ON SAME GRAPHS? by clicking on the YES button.

u. Repeat steps d through g using $N = 10$ and the following item parameters:

Item	**Difficulty**	**Discrimination**
1	b = -1.1	a = 1.9
2	b = -1.0	a = 1.7
3	b = -1.1	a = 1.8
4	b = -1.2	a = 1.6
5	b = -1.0	a = 1.9
6	b = - .8	a = 1.8
7	b = - .9	a = 1.9
8	b = -1.0	a = 1.9
9	b = -.9	a = 1.7
10	b = -1.0	a = 1.6

v. When the test characteristic curve appears, compare it to the previous curve that is still on the screen. Determine if you have increased the slope of the curve at the ability level -1.0.

w. When the test information function appears, compare it to the existing function on the screen. Determine if the maximum amount of information is larger than it was at an ability level of -1.0.

x. If all went well, the new set of test items should have improved the technical characteristics of the test as reflected in the test characteristic curve and the test information function.

Exercises

In each of the following exercises, establish a set of item parameters. After you have seen the test characteristic curve and the test information function, use the editing feature to change selected item parameter values. Also overlay the new curves on the previous curves. These procedures will allow you to see the impact of the changes. Repeat this process until you feel that you have achieved the test specification goal.

Exercise 1

Construct a ten-item screening test to function at an ability level of +.75 using a Rasch model.

Exercise 2

Construct a broad-range test under a three-parameter model that will have a horizontal test information function over the ability range of -1.0 to +1.0.

Exercise 3

Construct a test having a test characteristic curve with a rather small slope and a test information function that has a moderately rounded appearance. Use either a two- or three-parameter model.

Exercise 4

Construct a test that will have a nearly linear test characteristic curve whose mid-true score occurs at an ability level of zero. Use a Rasch model.

Exercise 5

Repeat the previous problem using a three-parameter model.

Exercise 6

Construct a test that will have a horizontal test information function over the ability range of -2.0 to +2.0, having a maximum amount of information of 2.5.

Exercise 7

Use the computer session to experiment with different combinations of testing goals, item characteristic curve models, and numbers of items. The goal is to be able to obtain test characteristic curves and test information functions that are optimal for the testing goals. It will be helpful to use the editing feature to change specific item parameter values rather than re-enter a complete set of item parameter values for each trial.

Things To Notice

1. Screening tests.

    a. The desired test characteristic curve has the mid-true score at the specified cutoff ability level. The curve should be as steep as possible at that ability level.

    b. The test information function should be peaked, with its maximum at the cutoff ability level.

    c. The values of the item difficulty parameters should be clustered as closely as possible around the cutoff ability of interest. The optimal case is where all item difficulties are at the cutoff point and the item discriminations are large. However, this is unrealistic because an item pool rarely contains enough items with common difficulty values. If a choice among items must be made, select items that yield the maximum amount of information at the cutoff point.

2. Broad-range tests.

    a. The desired test characteristic curve has its mid-true score at an ability level corresponding to the midpoint of the range of ability of interest. Most often this is an ability level of zero. The test characteristic curve should be linear for most of its range.

    b. The desired test information function is horizontal over the widest possible range. The maximum amount of information should be as large as possible.

    c. The values of the item difficulty parameters should be spread uniformly over the ability scale and as widely as practical. There is a conflict between the goals of a maximum amount of information and a horizontal test information function. To achieve a horizontal test information function, items with low to moderate discrimination that have a U-shaped distribution of item difficulties are needed. However, such items yield a

rather low general amount of information, and the overall precision will be low.

3. Peaked tests.

   a. The desired test characteristic curve has its mid-true score at an ability level in the middle of the ability range of interest. The curve should have a moderate slope at that ability level.

   b. The desired test information function should have its maximum at the same ability level as the mid-true score of the test characteristic curve. The test information function should be rounded in appearance over the ability range of most interest.

   c. The item difficulties should be clustered around the midpoint of the ability range of interest, but not as tightly as in the case of a screening test. The values of the discrimination parameters should be as large as practical. Items whose difficulties are within the ability range of interest should have larger values of the discrimination than items whose difficulties are outside this range.

4. Role of item characteristic curve models.

   a. Due to the value of the discrimination parameters being fixed at 1.0, the Rasch model has a limit placed upon the maximum amount of information that can be obtained. The maximum amount of item information is .25 since $P_i(\hat{\theta}) Q_i(\hat{\theta}) = .25$ when $P_i(\hat{\theta}) = .5$. Thus, the theoretical maximum amount of information for a test under the Rasch model is .25 times the number of items.

   b. Due to the presence of the guessing parameter, the three-parameter model will yield a more linear test characteristic curve and a test information function with a lower general level than under a two-parameter model with the same set of difficulty and discrimination parameters. The information function under a two-parameter model is the upper bound for

the information function under a three-parameter model when the values of *b* and *a* are the same.

c. For test specification purposes, the author prefers the two-parameter model.

5. Role of the number of items.

a. Increasing the number of items has little impact upon the general form of the test characteristic curve if the distribution of the sets of item parameters remains the same.

b. Increasing the number of items in a test has a significant impact upon the general level of the test information function. The optimal situation is a large number of items having high values of the discrimination parameter and a distribution of item difficulties consistent with the testing goals.

c. The manner in which the values of the item parameters are paired is an important consideration. For example, choosing a high value of the discrimination index for an item whose difficulty is not of interest does little in terms of the test information function or the slope of the test characteristic curve. Thus, the test constructor must visualize both what the item characteristic curve and the item information function look like in order to ascertain the item's contribution to the test characteristic curve and to the test information function.

References

Birnbaum, A. "Some latent trait models and their use in inferring an examinee's ability." Part 5 in F.M. Lord and M.R. Novick. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968.

Hambleton, R.K., and Swaminathan, H. *Item Response Theory: Principles and Applications.* Hingham, MA: Kluwer, Nijhoff, 1984.

Hulin, C. L., Drasgow, F., and Parsons, C.K. *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dow-Jones, Irwin: 1983.

Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum, 1980.

Mislevy, R.J., and Bock, R.D. *PC-BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models*. Mooresville, IN: Scientific Software, Inc, 1986.

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 Stone, M.A. *Best Test Design.* Chicago: MESA Press, 1979.

> **NOTE:** See (www.assess.com) for more information about the BILOG program. See (www.winsteps.com) for more information about BICAL and its successors, WINSTEPS and BIGSTEPS.

Selected Resources on Item Response Theory

Web Sites and Online Resources

Cumulative Item Response Theory Models
http://www.education.umd.edu/Depts/EDMS/tutorials/CIRT.html

Currently under development by the University of Maryland's Department of Measurement, Statistics, and Evaluation, this Web site will provide IRT models for binary and polytomous responses. Parametric and nonparametric models and applications will be addressed.

ERIC Clearinghouse on Assessment and Evaluation
http://ericae.net

The ERIC Clearinghouse on Assessment and Evaluation, hosted by the University of Maryland and funded by the U.S. Department of Education, maintains an extensive body of resources on the Web, including an online refereed journal called *Practical Assessment, Research & Evaluation* and a full-text library with more than 400 key reports from respected organizations.

Institute for Objective Measurement
www.rasch.org

The host of this site draws a careful distinction between IRT and Rasch measurement, but those interested in IRT will likely find items of interest here, including the full text of *Rasch Measurement Transactions*, the quarterly publication of the Rasch Measurement SIG of the American Educational Research Association.

Interactive CAT & IRT Mini-Tutorial
http://ericae.net/scripts/cat/catdemo.htm

Part of an online, interactive tutorial on computer adaptive testing developed, this mini-tutorial introduces the three-parameter IRT model and allows users to experiment with varying the item parameter values and generating graphs of item response functions.

Item Response Theory Models for Unfolding
http://www.education.umd.edu/EDMS/tutorials/Intro.html

This Web site, maintained by James S. Roberts, Assistant Professor in the Department of Measurement, Statistics and Evaluation at the University of Maryland, explores unfolding models for predicting item scores where responses are obtained in a rating scale format such as a Thurstone or Likert scale. Users can download free modeling software and illustrative data sets on attitudes toward capital punishment or censorship.

Electronic Discussion Groups

Rasch Discussion Listserv

This unmoderated forum sponsored by the Australian Council for Educational Research has operated since 1966 to support "the exchange of news, questions and answers about the theory and practice of Rasch Measurement."

To join, send an e-mail with text *subscribe rasch* to mailserv@acer.edu.au.

ERIC Database Search

Abstracts for several hundred documents and journal articles pertaining to item response theory have been indexed in the ERIC database. You can conduct your own search of ERIC for these citations via the Internet by using our online thesaurus-driven search engine, which we call the ERIC Search Wizard. Go to:

http://ericae.net/
Choose "Search ERIC"
You'll be taken to http://searcheric.org
Enter the term "item response theory" into the search box.

On the right side of the screen, you'll see a definition of the term and a list of related terms. Start your search by clicking on the box in front of Item Response Theory and then scroll down to the bottom of the column, where you will click on "Add to Set 1."

Use the related terms or the Look Up box to enter additional terms to combine with Item Response Theory in order to narrow your search. You may, for example, try Test Construction, Test Items, Item Bias, Item Analysis, Adaptive Testing, Test Reliability, or Test Validity in Set 2 or Set 3.

Print Classics

Applications of Item Response Theory to Practical Testing Problems
F. M. Lord
Lawrence Erlbaum, 1980

This classic text offers a thorough technical presentation of IRT models, including their limitations, as well as discussion of such practical problems as estimating ability and item parameters, equating, study of item bias, omitted responses and formula scoring. Flexilevel tests, multilevel tests, tailored testing, and mastery testing are also addressed.

Fundamentals of Item Response Theory
R. K. Hambleton, H. Swaminathan, and H. J. Rogers
Sage, 1991

This introductory text draws on concepts from classical measurement methods and basic statistics to present the basics of IRT and its applications in test construction, identification of potentially biased test items, test equating, and computerized adaptive testing. Alternative procedures for estimating IRT parameters, including maximum likelihood estimation, marginal maximum likelihood estimation, and Bayesian estimation, are discussed. Step-by-step numerical examples are included throughout.

Item Response Theory: Parameter Estimation Techniques
F.B. Baker
Marcell Dekker, 1992

This book presents the mathematical details of the parameter estimation procedures used in item response theory. The procedures maximum likelihood, marginal maximum likelihood, and Bayesian estimation are presented for binary, graded, and nominal response items. BASIC computer programs for these procedures are provided in the book.

Handbook of Modern Item Response Theory
W. J. van der Linden and R. K. Hambleton, eds.
Springer, 1997

This reference work provides an introduction to item response theory and its application to educational and psychological testing. A comprehensive treatment of models and families of models is provided in 27 chapters, each of which is authored by person(s) who either proposed or contributed substantially to the development of the model discussed. Each chapter includes an introduction, presentation of the model, parameter estimation and goodness of fit, and a brief empirical example. Some chapters also offer discussion.

Index