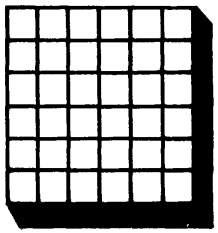

CHAPTER 8
Specifying the Characteristics of a Test

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:

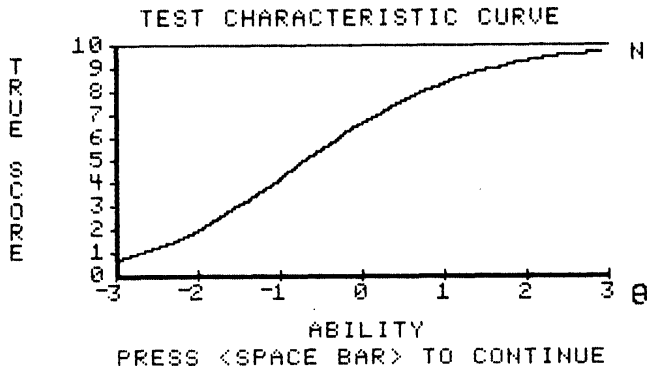

FIGURE 8-1. Test characteristic curve for the example

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.

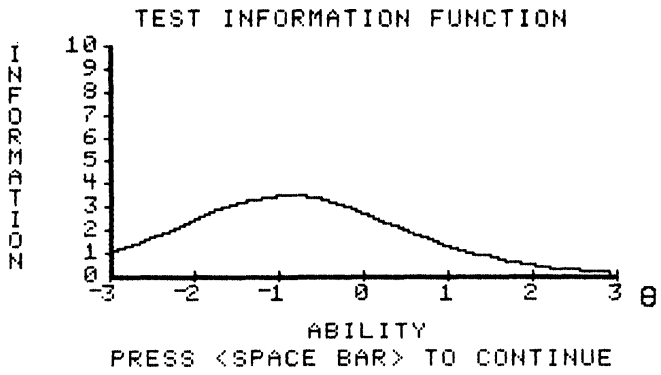

FIGURE 8-2. Test information function for the example

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 $P_i(\theta)) Q_i(\theta))$= .25 when $P_i(\theta)$ = .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.