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 θ_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 θ_j, the observed proportion of correct response is $p(\theta_j) = r_j/m_j$, which is an estimate of the probability of correct response at that ability level. Now the value of r_j can be obtained and $p(\theta_j)$ 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.

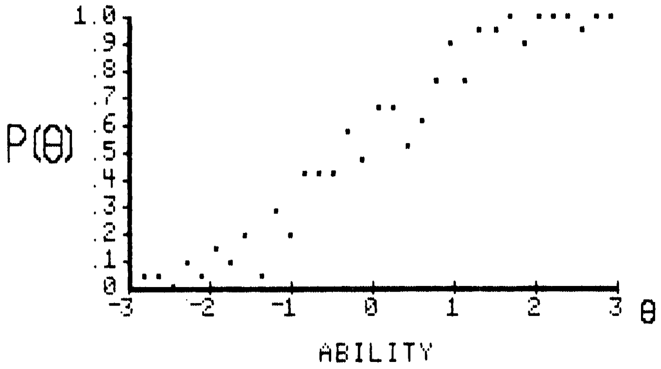

FIGURE 3-1. Observed proportion of correct response as a function of ability

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(θ_j) is computed at each ability level via the equation for the item characteristic curve model. The agreement of the observed value of p(θ_j) and computed value P(θ_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 $P(\theta_j)$ 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.

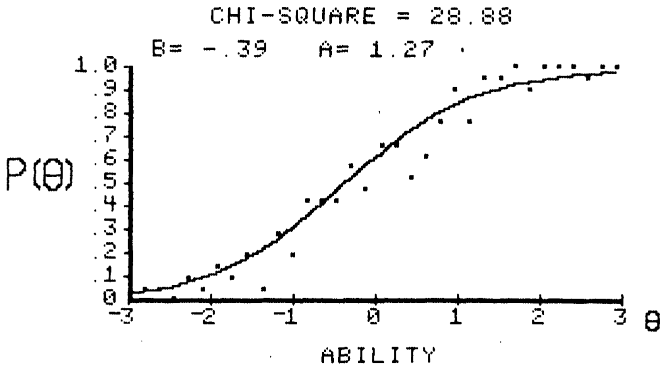

FIGURE 3-2. Item characteristic curve fitted to observed proportions of correct response

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.
θ_j is the ability level of group j.
m_j is the number of examinees having ability θ_j.
$p(\theta_j)$ is the observed proportion of correct response for group j.
$P(\theta_j)$ 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.

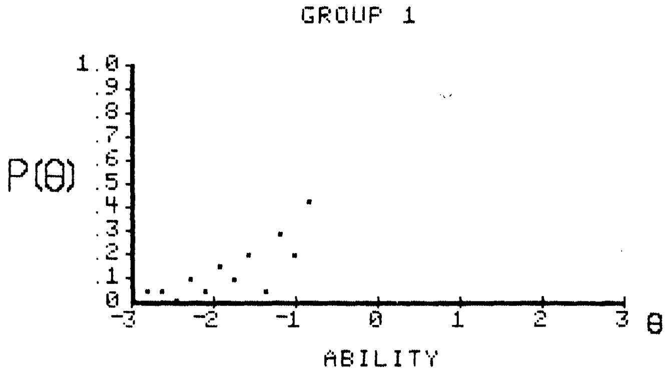

FIGURE 3-3. Observed proportions of correct response for group 1

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.

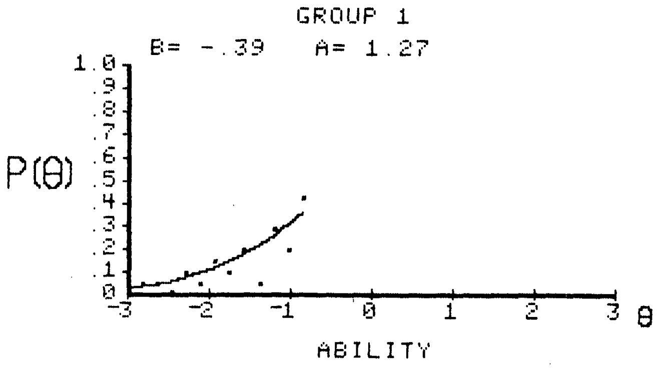

FIGURE 3-4. Item characteristic curve fitted to the group 1 data

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.

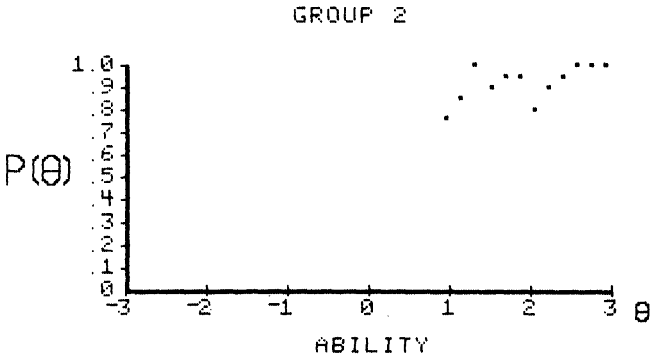

FIGURE 3-5. Observed proportions of correct response for group 2

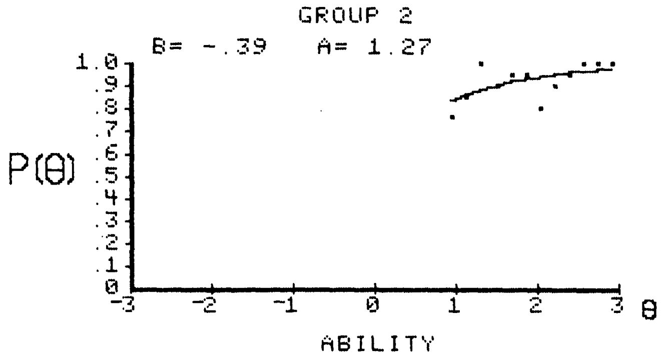

FIGURE 3-6. Item characteristic curve fitted to the group 2 data

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.

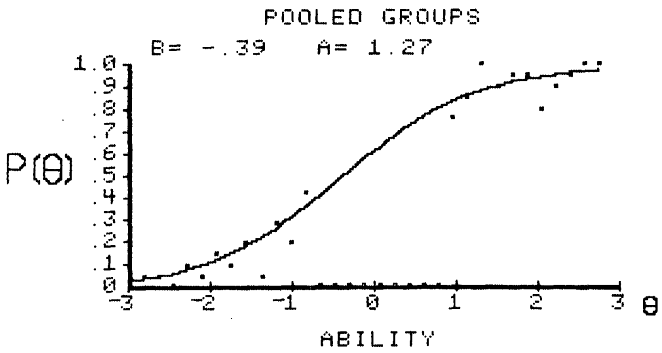

FIGURE 3-7. Item characteristic curve fitted to the pooled data, *b* = -.39 and *a* = 1.27

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 *UB* = +1

Group 2

LB = -1 *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 *UB* = -1

Group 2

LB = +1 *UB* = +3

and study the resulting display screens.

Then try:

Group 1

LB = -2 *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.