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.

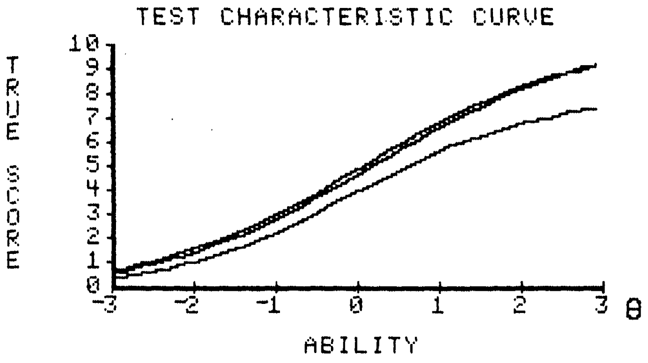

FIGURE 7-1. Test characteristic curves for the three data sets

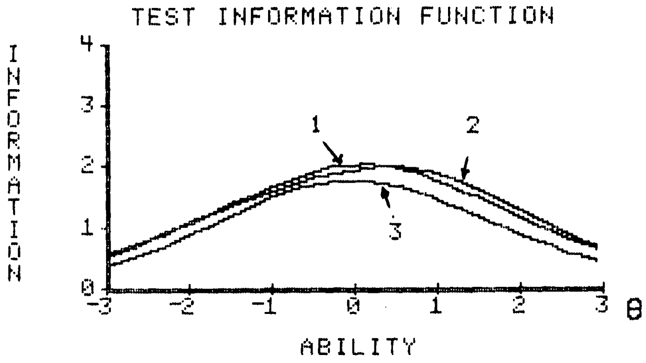

FIGURE 7-2. Test information function for the three data sets

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.