4

Building a Set of Items for Measurement

It would be a useful reference point to consider that there exists a family of Rasch models for measurement (Andrich, 1988; Masters & Wright, 1984). In this chapter we look at the use of the simplest model, the model for analyzing dichotomous data. This was, in fact, the model on which Georg Rasch did his initial work. Since then, the procedures for performing dichotomous Rasch analysis have been developed further by a number of researchers (foremost among them, Ben Wright from Chicago; Wright & Stone, 1979), whereas others have extended the basic Rasch model to include analysis of Likert-type rating scales (David Andrich, Perth, Australia, 1978a, 1978b, 1978c), responses that could be given partial credit (Geoff Masters, Melbourne, Australia, 1982), and testing situations in which many facets other than just person and item needed to be measured (Mike Linacre, Chicago, 1989). In each case, many researchers could claim that they contributed to the development of the Rasch-family models. However, the researchers that have been cited certainly are the most energetic proponents of these particular models. Each of these models from the Rasch family is more complex than the preceding one, but has the basic dichotomous model at its core. Therefore, if the researcher chooses not to use an added feature of any more complex model, it just collapses to the preceding, simpler model.

At the basis of all Rasch modeling is the model developed first: the model for analyzing dichotomous data, which are data that have simply two values, usually 0 and 1. It is easy to mistake this level of data as being "nominal," the sort of data we get when we categorize hair color as being brunette or blonde, or when we categorize a subject's sex as male or female. However, there is an important distinction concerning the data that is appropriate for analysis with the Rasch dichotomous model: The value of 1 is meaningfully greater than the value of 0, not merely different from 0. This might sound pedantic, but it is a very important point. If we allocate the code of 0 for the females in a sample, and 1 for the males, we intend just to differentiate them in terms of sex, showing that the sex of one group of respondents is different from that of the other group. However, when we use the code 1 to indicate the correct answer to a math problem and 0 as the code for the incorrect answer, we are saying something very different: Not only is the correct answer different from the incorrect answer; it also is better than the incorrect answer in a fundamentally important way. We regard the correct answer as superior to the incorrect answer, and we routinely regard children who get the correct answer as showing more ability than those who do not. Note then that Rasch modeling is appropriate only when we can impute some order in the allocation of scores such that 1 is better than 0 (as is a correct response versus an incorrect response). Order does not apply to the case in which 1 (e.g., male) is merely different from 0 (e.g., female), but certainly not better.

Two important points should be mentioned here. The first point is that the codes 1 and 0 merely record our observation of what the child actually did or did not do in response to the test prompt, not what the child could or could not do. Although we all might try to make measures out of the performances that we actually observe and record, and we might do this just so we can make decisions about the people who make the performances, we do not have any magic insight into how competent each person really is. All we have recorded is what the child did or did not do in response to the test prompt, not what the child could or could not do if given another response opportunity.

The second point is that researchers can save themselves a bit of hair tearing in the future by remembering always to use the code of 0 to record the lowest level of performance on any test item. Although this is obvious in the 0 = wrong and 1 = right format, it is not as obvious in coding a rating scale or awarding part marks. With Rasch analysis, it is a convenient and common practice to allocate 0 to indicate the lowest level of response and 1 the next level above that and so on. One routinely used format for the collection of dichotomous data is the multiple-choice test. Such a test would have only one "completely correct" or "best" answer that would receive the score of 1 for that item, with all the other distractors or alternative answers receiving the score of 0, although a partial-credit scoring and model might also be arranged from multiple-choice data (see chapter 7).

ANALYZING DICHOTOMOUS DATA: THE BLOT

In keeping with an important premise of this volume, that the key worked examples will be derived from the research of developmentalists, educators, and others trying to solve actual measurement problems, the dichotomous data discussed in this chapter come from a test of cognitive development for adolescents: Bond's Logical Operations Test (BLOT; Bond, 1976/1995). The BLOT was developed to provide a test suitable for administration to whole-class groups at a time, as a partial replacement for the individual interview technique developed and used by Jean Piaget and his colleagues in Geneva. The idea was to develop a multiple-choice test with response sheets that could be computer scored, so that a child's cognitive development could be categorized as more or less developed according to the total number of test items the child answered correctly. Of course, this general principle applies to most educational and psychological tests, so the principles outlined in the following discussion have far wider application than just to those interested in Piaget's idea of formal operational thinking.

One theme reiterated throughout this volume is that good tests have, as their basis, a very clear and explicit understanding concerning the line of inquiry the test is trying to put into practice—what was once termed "construct validity." Of course, this understanding might be revealed in a number of different ways. It could be part of a general psychological theory explained in one or more textbooks by some renowned guru, or part of a treatise on the exact sequence of development during a certain period of life. It might derive from a set of curriculum statements in a particular subject area at the grade school or high school level, or it might just as easily be taken from detailed theoretical or conceptual analysis of a field of knowledge being tested (e.g., math or spelling). In medical settings, it might be the understandings of rehabilitation progress after stroke, gleaned by medical professionals who reflect on the effects of their practice.

In the case of the BLOT, the specifications for the items were taken one by one from chapter 17 of the textbook entitled *The Growth of Logical Thinking* (Inhelder & Piaget, 1958). In this chapter Piaget spelled out in detail each of the logical operations that he thought were central to mature thought. The test developer's task then was to represent each of these logical specifications as accurately as possible in multiple-choice test items that would make sense to preadolescents and adolescents without requiring any specific background knowledge. As can be imagined, some items were rewritten a number of times as a result of trial runs with high school students.

Here, the key role of the test developer in putting the substantive theory into measurement practice is clear. In this case, it might have been handy to have Professor Piaget write the items, but then he was not interested in this aspect of group assessment at all. In all test development, the success of the enterprise will be determined largely by how well the intentions of the theory writer, the classroom teacher, or the medical specialist have been converted into items, not merely any items, but items such that the performances of the target audience will reveal exactly those intentions and not some other sort of ability. Clearly then, the test developer needs some detailed understanding of the substantive area of inquiry as well as a great deal of commitment to the implementing of that understanding into measurement practice.

The BLOT is a 35-item multiple-choice test that operationalizes item-by-item each of the schemas of the formal operational stage identified by Inhelder and

Piaget (1958). Each item comprises an item stem of two to four short sentences followed by a set of four or five alternative responses. The students' responses are collected on computer scan sheets and computer scored. The following interpretation shows us the sense that Rasch modeling can make of the BLOT and allows us to determine how much faith we can place in the idea that adolescents' cognitive development can be represented by the total raw score on the BLOT.

When using the BLOT, we generate a data file that looks like the following sample:

```
11111111110110101101011111111011111
11111111111111111111111111101111111
11010111111110111110111111010111111
11111111111111111111101111111111111
11111111111101111111011111111111111
11111111111110111101011111111111111
11111111111101111111011111111111111
11111111111111111111111111101011111
11111111111111111111111110111111111
11011111011111011111011111000110111
11111101111111111111011011111101111
11111101111111111111111111101001111
11111111111110111101011110111111111
11111111111101111101111111111111111
11111111111101111101111111111111111
11111111111101111111011111101110111
etc.
```

Each row represents the performances of one student on the 35 BLOT items. Given the principle of dichotomous scoring, the 1's represent the correct answers and the 0's represent the incorrect answers: The score for item 1 is in column 1, for item 2 in column 2, and so on up to item 35 in column 35. With this example, there is no student ID. The file is set up in the input order of the students' results. Although the BLOT can be computer scored, this file was typed in as a Word (text-only) file by the investigator.

ITEM DIFFICULTY LOCATIONS AND ERRORS

For the first part of the interpretation, we have included the results of the item analysis only, as Fig. 4.1. This is in exactly the same format as that described for the developmental pathway analogy introduced in chapter 3: easy items at the bottom and difficult items at the top. The error of the item difficulty estimate is shown by the comparative size of the item circle, whereas items that fit the Rasch model are located between the parallel dotted lines.

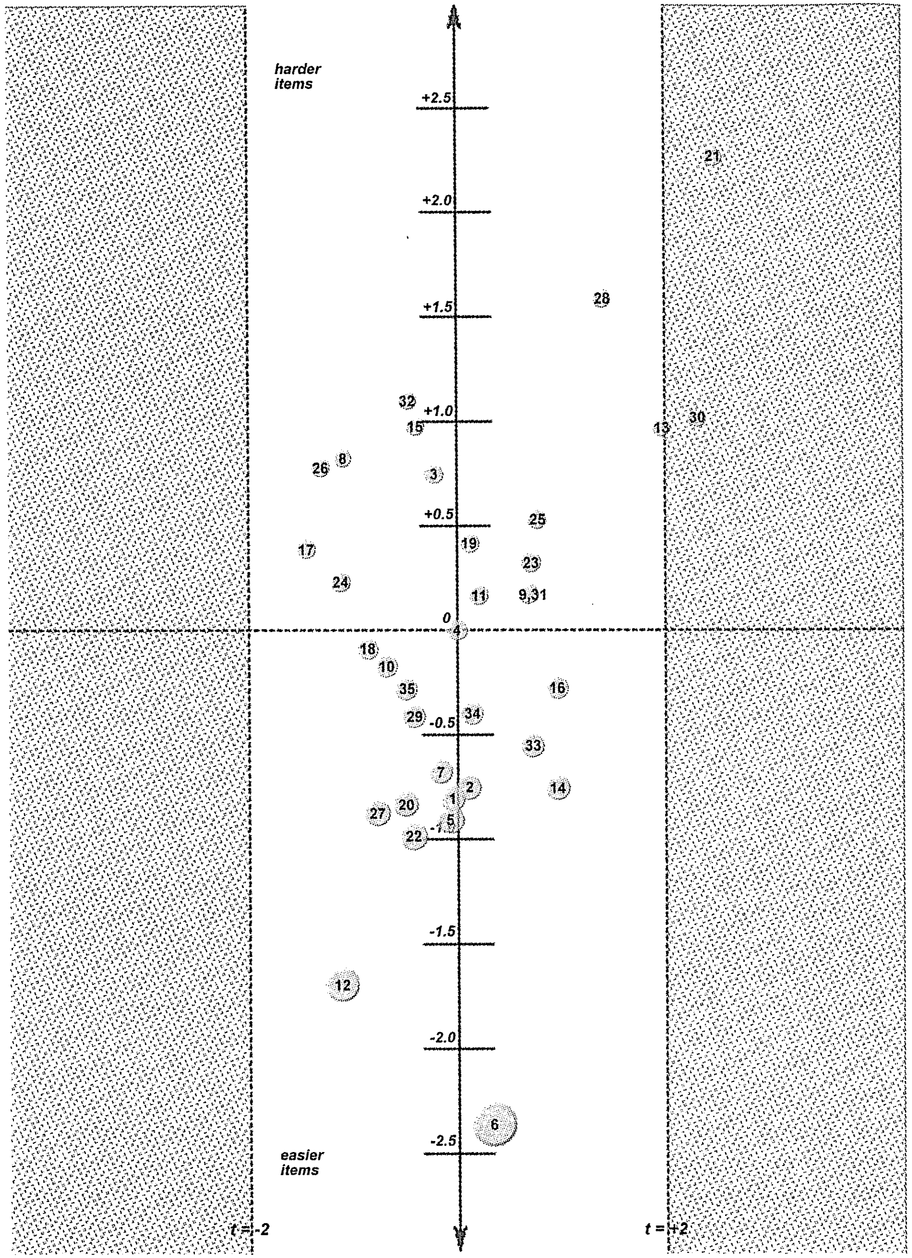

FIG. 4.1. BLOT item pathway.

A number of things can be immediately appreciated as a result of trying to find meaning in Fig. 4.1. First, it seems to represent some sort of developmental acquisition of ability: There are easy items, not-so-easy items, more difficult items, and even more difficult items. For example, items 6 and 12 are the very easiest BLOT items, and items 21 and 28 are extremely difficult in comparison with the others, whereas item 4 sits exactly at the midpoint (0) on the item difficulty scale. Given the range of varying difficulties of BLOT items, we might reasonably expect that a group of suitable students would show a range of developmental abilities on this test. It is worth noting that the extremely easy items (6 and 12) have the least precise estimates, whereas the error estimates for the remaining 33 items are comparatively quite small.

A glance at the dotted lines reveals that the fit of the BLOT to the Rasch model's expectations is pretty good. Locations for just two of the items (i.e., items 21 and 30) do not seem to fit satisfactorily to the same developmental pathway as do the remaining items. Items 21 and 30, therefore, should be candidates for closer inspection before they are included routinely in students' BLOT scores in the future. This is good evidence for reasonably inferring that the ability underlying the BLOT items follows a single line of inquiry. The Piagetian conception of cognitive development seems to be a reasonable description of that line of inquiry, given its explicit use in the BLOT development phase.

Although the item difficulties span five complete units on the logit scale, Fig. 4.1 shows that more than two logits of that development are represented by merely four items: 6 and 12 at the bottom of the scale and 21 and 28 at the top. However, from below −1 logits to above +1 logits, we have approximately 30 closely packed and overlapping items. The consequence of this is that we would find it very hard to locate persons precisely at the extreme ends of the scale represented by the BLOT items, but we could have a great deal of confidence if we had to make important decisions relating to students who perform in the −1 to +1 logits zone.

Although it is rather artificial to consider item performance separately from person performance, the purpose of this chapter is to demonstrate the development of a dichotomous test. At this point, suffice it to say that the distribution of person abilities among children who have been given the BLOT follows the same general pattern as that for items. The vast majority of person performances fit the Rasch model, whereas the distribution of persons along the ability scale is not as clumped as for items. This latter observation can be confirmed by referring to Fig. 4.2: the item–person map for the BLOT analysis.

ITEM FIT

Table 4.1 includes the item statistics from a Rasch analysis of dichotomous BLOT data. For each item number, the estimate of item difficulty and its accompanying error estimate in logits are given. These should correspond in a one-to-

```
--------------------------------------------------------------------------------------
BLOT Person Ability Estimates        BLOT Item Difficulty Estimates
--------------------------------------------------------------------------------------
  4.0
                   XXXXXXXXXXXXXXX  |
                                    |
                          XXXXXXXX  |
  3.0                               |
                           XXXXXXX  |
                                    |
               XXXXXXXXXXXXXXXXXXX  |      21
                            XXXXXX  |
  2.0                               |
                           XXXXXXX  |
                                    |
                             XXXXX  |      28
                         XXXXXXXXX  |
                                    |
                      XXXXXXXXXXXX  |      32
                       XXXXXXXXXXX  |      30
  1.0                        XXXXX  |      13      15
                          XXXXXXXX  |       8      26
                                    |       3
                               XXX  |      25
                               XXX  |      17      19
                          XXXXXXXX  |      23
                            XXXXXX  |       9      11      24      31
   .0                           XX  |       4
                                XX  |      18
                                 X  |      10
                                 X  |      16      35
                                 X  |      29      34
                                 X  |       7      33
                               XXX  |       1       2      14
                                    |      20
 -1.0                            X  |       5      27
                                 X  |      22
                                    |
                                 X  |
                                    |
                                    |      12
 -2.0                               |
                                 X  |
                                    |
                                    |       6
                                    |
 -3.0                               |
--------------------------------------------------------------------------------------
  Each X represents     1 student
```

FIG. 4.2. Item–person map for the BLOT analysis (QUEST).

one way with the pictorial representation in the earlier figure, although only some of the BLOT items are numbered: the higher the difficulty estimate, the further up the pathway, and the larger the error estimate, the larger the item stepping-stone. However, the columns that contain the fit statistics are not so easily interpreted.

Generally speaking, fit statistics focus on two aspects of fit, each of which is routinely reported in both an unstandardized and a standardized form. In Table 4.1 the two aspects of fit reported are item infit and outfit. The unstandardized form is reported as mean squares, and the standardized form is reported as a *t* statistic, in which acceptable values are those routinely accepted for *t* (i.e., –2 to +2). The mean square is the unstandardized form of the fit statistic and merely the mean, or average value, of the squared residuals for that item. The residual values represent the differences between the Rasch model's theoretical expectation of item performance and the performance actually encountered for that item in the data matrix. Larger residuals mean an item with larger differences between how the item should have performed (i.e., Rasch model expectations) and how it actually performed. Residuals are squared, following the usual statistical convention, to make all "actual minus expected" differences positive so they can be added to give a sum of differences. The concept of fit is the subject of chapter 12.

In the standardized versions of fit statistics, the mean square value is transformed, with the sample size kept in mind, to produce a statistic with a distribution just like *t*. The "fit" issue will be raised again and again in this volume and everywhere that Rasch analysts gather to chat (e.g., see Smith, 2000).

The infit and outfit statistics adopt slightly different techniques for assessing an item's fit to the Rasch model. The infit statistic gives relatively more weight to the performances of persons closer to the item value. The argument is that persons whose ability is close to the item's difficulty should give a more sensitive insight into that item's performance. The outfit statistic is not weighted, and therefore is more sensitive to the influence of outlying scores. It is for this reason that users of the Rasch model routinely pay more attention to infit scores than to outfit scores. Aberrant infit scores usually cause more concern than large outfit statistics. Of course, outfit statistics do have meaning, and we return to the issues involved in interpreting infit and outfit statistics in chapter 12.

INTERPRETING RASCH ANALYSIS OUTPUT

To make this analysis work, you would need to tell your Rasch software:

The name of the data file and where it is located.

The format of the data: easy in this case, 35 items (one item per column usually is the default).

TABLE 4.1
BLOT Item Difficulty Estimates With
Associated Error Estimates for Each Item

Item	*Difficulty Estimate*	*Error Estimate*	*Infit Mean Square*	*Outfit Mean Square*	*Infit t*	*Outfit t*
1	−0.77	0.26	0.98	0.69	0.0	−0.8
2	−0.70	0.26	1.01	0.75	0.1	−0.6
3	0.74	0.2	0.98	0.9	−0.2	−0.5
4	0.00	0.22	1.00	0.88	0.0	−0.4
5	−0.98	0.28	0.98	0.76	−0.1	−0.5
6	−2.42	0.47	1.06	0.83	0.3	0.1
7	−0.64	0.25	0.97	0.65	−0.1	−1.0
8	0.85	0.19	0.91	1.00	−1.1	0.1
9	0.18	0.21	1.07	0.97	0.7	0.0
10	−0.19	0.23	0.92	0.68	−0.7	−1.1
11	0.18	0.21	1.02	0.96	0.2	−0.1
12	−1.76	0.36	0.69	0.24	−1.1	−1.5
13	1.00	0.19	1.16	1.32	2.0	1.8
14	−0.70	0.26	1.15	1.32	1.0	0.9
15	1.00	0.19	0.96	0.84	−0.4	−0.9
16	−0.30	0.23	1.13	1.03	1.0	0.2
17	0.39	0.2	0.87	0.75	−1.4	−1.2
18	−0.05	0.22	0.9	0.74	−0.9	−1.0
19	0.47	0.2	1.01	1.05	0.1	0.3
20	−0.84	0.27	0.91	0.81	−0.5	−0.4
21	2.33	0.2	1.27	1.75	2.6	3.4
22	−1.06	0.29	0.91	1.69	−0.4	1.4
23	0.35	0.21	1.06	0.92	0.7	−0.3
24	0.22	0.21	0.89	1.03	−1.1	0.2
25	0.51	0.2	1.07	1.26	0.8	1.2
26	0.78	0.2	0.89	0.75	−1.3	−1.4
27	−0.91	0.27	0.85	0.62	−0.8	−0.9
28	1.63	0.19	1.12	1.23	1.4	1.4
29	−0.46	0.24	0.94	0.71	−0.4	−0.8
30	1.07	0.19	1.19	1.15	2.3	0.9
31	0.18	0.21	1.07	1.55	0.7	2.0
32	1.14	0.19	0.96	0.85	−0.5	−0.9
33	−0.52	0.25	1.1	0.93	0.7	−0.1
34	−0.41	0.24	1	0.79	0.1	−0.6
35	−0.30	0.23	0.93	0.73	−0.5	−0.9

Note. Fit statistics are shown in their natural (mean square) and standardized forms (standardized as *t*).

The type of analysis: easy again, dichotomous is the usual default.
The name and location for the output file.

Most versions of Rasch analysis software produce some form of the item–person map shown as Fig. 4.2, in which the items are indicated by the item number, and each individual person's performance is represented by an "X." The delightful thing about this Rasch representation of data analysis is that many of the person and item relations are shown in meaningful pictorial, or "map," form.

The logit scale, which is the measurement unit common to both person ability and item difficulty, is displayed down the middle of the map in Fig. 4.2. Because the logit scale is an interval scale, the equal distances anywhere up and down that scale have equal value. Therefore, item 15 is as much more difficult than item 4 as item 4 is more difficult than item 5. The distances between are equal (1 logit). Of course, the same equal-value principle applies to differences in person locations as well. Persons and items are located on the map according to their ability and difficulty estimates, respectively.

As a convenient starting point for the mapping process, the mean of the item difficulties is adopted by default as the 0 point. In this case, ignoring the error of measurement for a moment, item 4 is calculated as having that exact difficulty estimate (0 logits), so it is located at the 0 point on the item–person map. Person locations are plotted so that any person has a 50% probability of succeeding with an item located at the same point on the logit scale. For example, a person with an ability estimate of 0 logits has a 50% probability of succeeding on item 4. That same person would have a greater than 50% chance of succeeding on items less difficult than item 4 (say, items 18, 29, and 5) and a less than 50% probability of succeeding on items more difficult than item 4 (say, items 17, 25, and 26). The 50% *limen*, or threshold, is adopted routinely by Rasch analysis, although some Rasch software allows for variations from this value to be specified. For example, those committed to the concept of mastery learning might want to use the 80% threshold that is used routinely to assess mastery.

With those basic principles in mind, we now can tell immediately from the item–person map in Fig. 4.2 that the BLOT is too easy for a sample like this one. Just look where the persons are located in comparison with the items. First, the person distribution is top-heavy in comparison with the item distribution. Second, the top 50 BLOT performers (one third of this sample) are targeted by only two questions: items 21 and 28. The Rasch output tells us as well that three candidates topped out on the BLOT with a perfect score of 35 of 35. From a general test-development perspective, this would be regarded as a serious inadequacy in a test. If this is the usual sort of target group for this test, then the test needs some more questions of a difficulty like that of 21 and 28 so the abilities of the high-fliers can be more precisely estimated. Also, we would need some even more difficult questions to raise the "ceiling" of the test.

A key point to remember, however, is that Rasch analysis item–person maps usually report the relations between the two key variables only: item difficulty estimates and person ability estimates. Other key parts of the analysis—the precision of those estimates (error), the fit of the items, the fit of the persons, the reliabilities of the person and item estimates—are reported in detail in the output tables.

For items we have the following information that is useful:

```
Summary of Item Estimates
Mean                              0.00
SD                                0.95
SD (adjusted)                     0.92
Reliability of estimate           0.94

Fit Statistics
Infit Mean Square                    Outfit Mean Square
 Mean                  1.00           Mean     0.95
 SD                    0.11           SD       0.31

        Infit t                        Outfit t
 Mean                  0.09           Mean    -0.05
 SD                    0.98           SD       1.10

0 items with zero scores
0 items with perfect scores
```

We already know that the mean of item estimates is located at 0 (by default), and that the standard deviation for item estimates is nearly 1. We can confirm the latter by referring to the item–person map: The vast majority of items are located in the narrow band between +1 and −1 logits. The reliability of the item difficulty estimates is a very high .94 on a 0 to 1 scale. Item reliability can be interpreted on this 0 to 1 scale, much in the same way as Cronbach's alpha is interpreted, or it can be transformed to an item separation index, wherein the reliability is calculated as the number of standard errors of spread among the items (see Fox & Jones, 1998, or Wright & Masters, 1982, for an explanation). Item reliability and item separation refer to the ability of the test to define a distinction hierarchy of items along the measured variable. The higher the number, the more confidence we can place in the replicability of item placement across other samples. Therefore, the item reliability index of .94 means that we can quite readily rely on this order of item estimates to be replicated when we give the BLOT to other samples for whom it is suitable.

The summary of fit statistics also can be informative. Unstandardized fit estimates (i.e., mean squares) are modeled by the Rasch algorithm to have a mean of 1. The actual unstandardized item fit statistics for the BLOT have their means very close to the expected 1, with the infit mean squares showing little spread from that ideal and the outfit mean squares much greater variation.

In the standardization of fit scores, the mean square values are transformed so they are distributed like *t*, with a mean of 0 and a standard deviation of 1. Therefore, we should not be surprised to see the preceding raw item mean squares transformed into near-0 values. But for how many of the BLOT items is this information applicable? The little note at the bottom of the output reminds us that all the BLOT items were useful for this sample. An item would not be useful for discriminating ability among members of this group if everyone was successful with it (item too easy) or everyone got it wrong (item too hard).

COMPARING PERSONS AND ITEMS

When we turn our focus toward the summary of person performances, we find that Rasch modeling has the distinct advantage of applying the same analytical logic, and therefore the same logic of interpretation, to persons as it does to items.

```
Summary of Case Estimates
Mean                              1.56
SD                                1.30
SD (adjusted)                     1.17
Reliability of estimate           0.81

Fit Statistics
Infit Mean Square                    Outfit Mean Square
 Mean            0.99                 Mean     0.95
 SD              0.13                 SD       0.46

      Infit t                          Outfit t
 Mean            0.13                 Mean     0.10
 SD              0.58                 SD       0.63

0 cases with zero scores
3 cases with perfect scores
```

The person ability estimate mean of +1.56 is the first indicator that this sample finds this test comparatively easy. Figure 4.3 shows three possible relations between item difficulty and person ability. The mean person estimate (i.e., the group average) would be closer to 0 for a well-matched test (Fig. 4.3b). A tough test would yield a mean person estimate with a large negative value (Fig. 4.3c). The standard deviation of 1.30 for person estimates indicates greater spread of person measures or variation in those measures than with item measures. The reliability of the person ability estimates is high at .81, which is not as reliable as the item separations, but more than acceptable nonetheless.

This corroborates the targeting problem we identified from the item–person map. Although we can rely on this order of person estimates to be replicated when we give these persons another test like the BLOT, in the current analysis we have better information about the items than we do about the persons, so the item estimates are more reliable. In other words, the performances of 150 persons give us more good information about the 35 BLOT items than the 35 BLOT items give about the 150 persons. From consideration of the three distributions in the item–person maps of Fig. 4.3, we could expect the best person separability index in case b, where the match between items and persons is the best. In case c, the difficult test, both item and person reliability would be lower: The least able persons have no items to distinguish between them, whereas the toughest questions have no persons sufficiently able to provide good information about them.

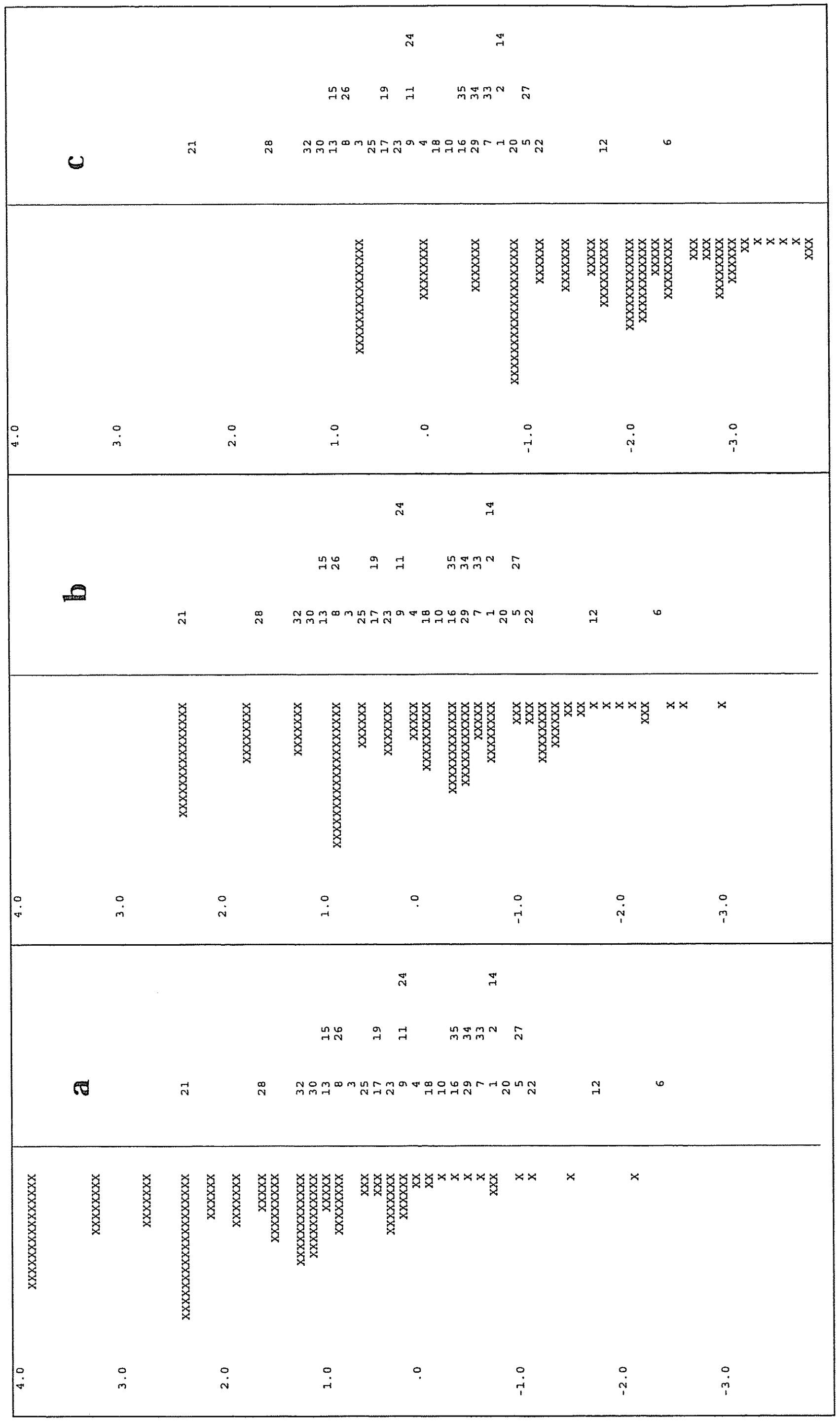

FIG. 4.3. Item–person maps showing the test as (a) relatively easy for the sample, (b) well-matched to the sample, and (c) relatively difficult for the sample.

Again, the person fit summary statistics are equally good. The mean of the infit mean squares at 0.99 and the outfit mean squares at 0.95 are very close to the Rasch-modeled expectations of 1. Consequently, they produce standardized fit t values just greater than 0. The spread in, or variation of, modeled fit scores for persons (infit t $SD = 0.58$ and outfit t $SD = 0.63$) suggests that the vast majority of person ability estimates will have error estimates well inside the conventionally acceptable range of −2 to +2.

THE THEORY–PRACTICE DIALOGUE

Of course, every test developer and user should try to discern what the results from the performance of the items and persons in practice have to say about the substantive theory being investigated, and should try to decide what the theory tells about the persons and items under investigation. This should always be seen as an ongoing dialectical process. We have included a little of it here to indicate the sort of meanings that might be attributed to the results of the analysis shown earlier.

The analysis provides pretty good evidence that the items work well together to represent one underlying path of inquiry or ability. Given that the specifications for the logical structure of each and every item were lifted directly from the Inhelder and Piaget (1958) text, this could be seen to confirm the idea that Piaget's model for adolescent intelligence is coherent in itself. At least psychometric evidence points to "something" and not "many things" as the object of inquiry. Moreover, whatever this ability is, it also is evident in the BLOT-answering behavior of a bunch of suitable subjects: 150 adolescent schoolchildren.

Because both the items and the persons were shown to behave in sufficiently lawful and predictable ways, it is reasonable to conclude that this part of Piaget's theory and the BLOT interpretation of it are certainly worth the effort of continued refinement and investigation.

The ceiling effect on the BLOT continues to be an ongoing problem: The most cognitively developed kids top out on the test. Although that amounts to only 3 of the 150 tested for this chapter (at age 15), we could reasonably expect that more and more of these students would "hit the ceiling" as we tracked their development over time. This is further complicated to the extent that some Rasch analysis software routinely imputes an ability estimate for those who get a perfect score, whereas other software packages ignore the perfect scorers because they do not have enough information to provide an accurate estimate. Clearly, the BLOT needs more difficult items based on Piaget's specifications if we intend to use it to estimate accurately the cognitive development of our more intellectually able teenagers.

The spread of items, or the lack of spread, on the item–person map suggests that some of the BLOT items are redundant: The area from 0 to −1 logits is saturated with items. It seems that a number of the particular intellectual skills incorporated into BLOT items are very much like other skills/items, and that it would not be necessary to include them all in a parsimonious test. Indeed, dropping some of the psychometrically redundant items in favor of more difficult items would remedy two of the apparent deficiencies of the BLOT.

Of course, psychometrically redundant and theoretically redundant are two different but related perspectives on the theory–practice nexus: In the first round, the practice tells us that the conceptualization of the theory has a lot going for it, but that a more useful test could be developed by going back to the theory to find specifications for further item development and rationalization.

Software control files for this analysis and their explanations appear as follows:

QUEST:

```
title BLOT for Chapter Four
data bond87.txt
format items 5-39
est
show>>BLOT.out
show items>>Blot.items
quit
```

Line 1 gives a name to the output.
Line 2 tells QUEST which file has the data.
Line 3 that indicates that the BLOT responses are in columns 5 to 35.
Line 4 commands QUEST to perform a Rasch *est*imation.
Line 5 directs the general output to a file called BLOT.out.
Line 6 directs the item statistics output to a file called Blot.items.

WINSTEPS:

```
&INST
TITLE='BLOT for Chapter Four'
NI=35
ITEM1=5
NAME1=1
IFILE=BLOT.IF
&END
Item 1
Item 2
Item 35
END NAMES
11111111110110101101011111111011111
11111111111111111111111111101111111
```



```
11010111111111011111011111101011111
etc.
```

Line 1 contains a command that must begin every WINSTEPS file.
Line 2 provides a title for the output.
Line 3 indicates the number of items in the test.
Line 4 identifies the starting column for the data.
Line 5 identifies the starting column for the person identification number.
Line 6 directs the item statistics output to a file called Blot.if.
Line 7 indicates the end of the commands and the beginning of the item names.
Lines 8 to 10 give a line-per-item name. Only the first two and the last BLOT items are named here.
Line 11 indicates an end to the names.
Line 12 etc. ASCII data file, like the 0's and 1's shown earlier in the chapter, follows immediately after this line.