Chapter 11
Equating—Linking Instruments Through Common Items

Linking of Instruments with Common Items

In many areas of social measurement, different instruments but with *some common items*, have been constructed to assess the same variable, and it is considered important to place them on the same scale. In these cases of *some common items* between two instruments, the implication is that not all persons have attempted the same items. We comment on applications of this feature after we describe a method for applying the Rasch model for analyzing a data matrix when not all persons have attempted all items. Often, when two sets of items with some common items are placed on the same scale, it is said that the sets of items have been *linked*. Such a design was in the original work of Rasch which led him to his theory of measurement (Rasch, 1960).

Linking Three Items Where One Item Is Common to Two Groups

To illustrate the procedure, we expand on the estimation of item locations in Chap. 9 where we estimated the relative difficulties of two items, items 1 and 8 from Table 3.1 in Chap. 3. We consider the case where item 8 and another item, called item 11 that assessed the same content and could have been in the same test, have been answered by another group of people for whom the test is relevant. Thus, item 8 is common to two groups of persons, and items 1 and 11 are answered by only one of the two groups. Figure 11.1 shows the design of the administration of these three items.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_11

Estimating Differences Between Difficulties and then Adjusting the Origin

In Chap. 9, we already estimated the difference between difficulties of items 1 and 8. We consider that this estimate has come from responses by group 1 in Fig. 11.1. The relative difficulties were $\hat{\delta}_1 = -0.973$, $\hat{\delta}_8 = 0.973$ where, because we can estimate only a difference, we made $\hat{\delta}_1 + \hat{\delta}_8 = 0$. We use the same procedure to estimate the difference between difficulties of items 8 and 11 as we did to estimate the difference between difficulties of items 1 and 8.

Table 11.1 shows the relevant responses to items 8 and 11 from group 2. It shows that seven persons had a total score of 1 on the two items.

Following the procedures of Eq. (9.4) in Chap. 9 we have

$$\text{Proportion}\{(x_{n8} = 1, x_{n11} = 0)|r_n = 1\} = \frac{5}{7} \text{ and}$$
$$\text{Proportion}\{(x_{n8} = 0, x_{n11} = 1)|r_n = 1\} = \frac{2}{7}.$$

Substituting these proportions as estimates of the respective probabilities in Eq. (9.5) of Chap. 9 gives

$$\ln\left[\frac{5/7}{2/7}\right] = \ln\left[\frac{5}{2}\right] = \ln[2.5] = \hat{\delta}_{11} - \hat{\delta}_8.$$

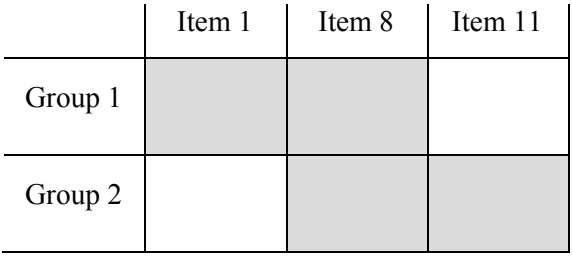

Fig. 11.1 Linking design for three items with one item common to two groups

Table 11.1 Responses of persons to items 8 and 11 given $\{r_n - 1\}$

Person	Person count	Item 8	Item 11	$\{r_n - 1\}$	a_{811}
3	1	1	0	1	1
4	2	1	0	1	1
5	3	1	0	1	1
6	4	1	0	1	1
7	5	0	1	1	0
8	6	0	1	1	0
9	7	1	0	1	1
	Total = 7				Sum = 5

That is,

$$\hat{\delta}_{11} - \hat{\delta}_8 = \ln[2.5] = 0.916.$$

Setting $\hat{\delta}_8 + \hat{\delta}_{11} = 0$, gives $\hat{\delta}_8 = -0.458$ and $\hat{\delta}_{11} = 0.458$.

Now we have two values for Item 8, $\hat{\delta}_8 = 0.973$ from the comparison with item 1 with the responses from group 1, and $\hat{\delta}_8 = -0.458$ from a comparison with item 11 obtained from group 2. To place estimates of all three items on the same scale we note that the origin is arbitrary in each set of estimates and that only the difference between item difficulties has been estimated.

Thus, we can add constants to the estimates providing we preserve the differences. We can simply retain the value of item 8 as estimated from group 1, find the difference with its value obtained from group 2, and then add the same value to item 11.

The difference between the two estimates for Item 8 is $0.973 - (-0.458) = 1.431$. Adding 1.431 to both the estimates of items 8 and 11 from group 2 gives $\hat{\delta}_8 = 0.973$ and $\hat{\delta}_{11} = 1.889$. Thus, now item 8 has the same estimate in group 2 as in group 1, and the difference of 0.916 between items 8 and 11 obtained from group 2 has been retained. Table 11.2 summarizes the calculations. In this calculation, the average of the difficulties of all three items is 0.630.

If it is deemed convenient for some reason that the sum of these item difficulties is 0, then this can be achieved simply by subtracting the average difficulty of the items from each item. The estimates with this subtraction, which give $\hat{\delta}_1 + \hat{\delta}_8 + \hat{\delta}_{11} = 0$, is shown in the last row of Table 11.2.

Table 11.2 shows that item 11 is more difficult than item 8 and very much more difficult than item 1. Perhaps group 2 was more proficient than group 1 and that is the reason that the more difficult item was given to this group.

The case of three items was shown above for purposes of exposition. In general, there are of course many items in each test, and more than one common item. The generalization of the procedure above, where there are many items, is to calculate the *mean of the common items* in the two groups, and then add the difference between these means to all items of one of the sets of items. Another procedure is to analyze all the responses of all the items and take advantage of the analysis which handles missing responses. In Fig. 11.1 responses of group 2 to item 1 and group 1 to item 11 are said to be missing. This procedure is described next.

Table 11.2 Estimates of items 1, 8 and 11 placed on the same scale

Items	$\hat{\delta}_1$	$\hat{\delta}_8$	$\hat{\delta}_{11}$	Mean
Group 1	−0.973	0.973		
Group 2		−0.458	0.458	
0.973 − (−0.458)		1.431	1.431	
Estimates	−0.973	0.973	1.889	0.630
Estimates Mean 0.0	−1.603	0.343	1.259	0.000

Estimating Differences Between Difficulties Simultaneously by Maximum Likelihood

We now summarize the approach that can estimate the parameters simultaneously in the case that not all persons respond to all items. We use the example of the three items 1, 8 and 11 with data from Table 11.1 of this chapter and Table 9.3 of Chap. 9. We show this because it is the basic method used by computer programs and we think it helps to understand the principles by which the programs provide estimates.

Before proceeding, we show how the complementary equations, Eqs. (9.1) and (9.2) of Chap. 9 with respect to items 1 and 8, can be written as a single equation.

These equations are

$$P_{1.8|1} = \Pr\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\} = \frac{e^{-\delta_1}}{e^{-\delta_1} + e^{-\delta_8}} \tag{11.1}$$

$$P_{8.1|1} = \Pr\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\} = \frac{e^{-\delta_8}}{e^{-\delta_1} + e^{-\delta_8}} \tag{11.2}$$

We introduce the simplified notation of $P_{i.j|1}$ because we use it below to help summarize the solution equations. The first subscript indicates the item which has the response 1, and the second the one that has the response 0.

Equations (11.1) and (11.2) can be written as a single equation in the form

$$\Pr\{(X_{n1} = x_{n1}, X_{n8} = x_{n8})|r_n = 1\} = \frac{e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{e^{-\delta_1} + e^{-\delta_8}}. \tag{11.3}$$

It is evident that when $\{(X_{n1} = 1, X_{n8} = 0)|r_n = 1\}$ and the values are substituted in Eq. (11.3), that it results in Eq. (11.1), and that when $\{(X_{n1} = 0, X_{n8} = 1)|r_n = 1\}$ and the values are substituted in Eq. (11.3), that it results in Eq. (11.2).

In general, for any two items i, j Eq. (11.3) generalizes to

$$\Pr\{(X_{ni} = x_{ni}, X_{nj} = x_{nj})|r_n = 1\} = \frac{e^{-x_{ni}\delta_i - x_{nj}\delta_j}}{e^{-\delta_i} + e^{-\delta_j}}. \tag{11.4}$$

From Eq. (11.3), and focusing on just the two items 1 and 8, we can write the *likelihood* L of the responses. There are 16 cases in Table 9.3 of Chap. 9 which have a total score of $r_n = 1$ and we can, therefore, write L of the set of responses as the product of these probabilities (which are conditional on a total score of 1); it is called a conditional likelihood. Thus

$$L = \prod_{n=1}^{16} \frac{e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{e^{-\delta_1} + e^{-\delta_8}} = \frac{\prod_{n=1}^{16} e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}}$$

$$= \frac{e^{-\sum_{n=1}^{16} x_{n1}\delta_1 - \sum_{n=1}^{16} x_{n8}\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}}$$

$$= \frac{e^{-\delta_1 \sum_{n=1}^{16} x_{n1} - \delta_8 \sum_{n=1}^{16} x_{n8}}}{(e^{-\delta_1} + e^{-\delta_8})^{16}}$$
$$= \frac{e^{-14\delta_1 - 2\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}} \tag{11.5}$$

The coefficients $\sum_{n=1}^{16} x_{n1}$ and $\sum_{n=1}^{16} x_{n8}$ of δ_1 and δ_8, respectively, are 14 and 2 (the sum of the responses), which is the number of times each one has a score of 1 when the other has a score of 0.

Taking the logarithm gives

$$\ln L = -14\delta_1 - 2\delta_8 - 16 \ln(e^{-\delta_1} + e^{-\delta_8}). \tag{11.6}$$

We need calculus to derive equations that give values of δ_1 and δ_8 that maximize the value of Eq. (11.6). There is one equation for each item. These are obtained by differentiating Eq. (11.6) first with respect to δ_1 and then with respect to δ_8.

This gives for the respective items

$$\delta_1 : -14 + 16\frac{e^{-\hat{\delta}_1}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = 0 \tag{11.7}$$

and

$$\delta_8 : -2 + 16\frac{e^{-\hat{\delta}_8}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = 0. \tag{11.8}$$

We have placed a "hat" on the parameters to indicate that when they satisfy these equations, they are estimates. It is also evident that the ratios that involve the parameters are simply the conditional probabilities of Eqs. (11.1) and (11.2). As a result, we can write

$$\delta_1 : -14 + 16\hat{P}_{1.8|1} = 0 \tag{11.9}$$

and

$$\delta_8 : -2 + 16\hat{P}_{8.1|1} = 0. \tag{11.10}$$

Equations (11.9) and (11.10) have the form of the solution for a binomial variable which can be seen by writing them as

$$\delta_1 : 16\,\hat{P}_{1.8|1} = 14 \tag{11.11}$$

and

$$\delta_8 : 16\hat{P}_{8.1|1} = 2. \tag{11.12}$$

However, these two equations are not independent, and therefore there are many solutions that satisfy them. One way of telling that the equations are not independent is to check if the sum of the equations reduces to an identity. In fact, we can see that, because $\hat{P}_{1.8|1} = 1 - \hat{P}_{8.1|1}$ the sum of the left side of the two equations, $16\hat{P}_{1.8|1} + 16\hat{P}_{8.1|1} = 16(1) = 16$, is exactly the sum of their right-hand sides, $14 + 2 = 16$. The dependence arises because, although there is a parameter for each item, the only information available is about their difference. To obtain a solution for Eqs. (11.11) and (11.12) that can be agreed upon, it is conventional to fix the sum of the estimates to be 0, as we did in Chap. 9. It is, however, possible to fix the value of one of the items and let the other take the estimate from the responses. Thus the additional equation generally specified is

$$\hat{\delta}_1 + \hat{\delta}_8 = 0. \tag{11.13}$$

The solution to these equations is found iteratively, as shown for person estimates in Chap. 10. That is, initial values are placed on the left side of Eqs. (11.11) and (11.12), and then based on their differences from 0, adjustments are made with the constraint of Eq. (11.13) imposed with each iteration until the difference is small enough to be acceptable, perhaps 0.0001.

We do not go through the process, but for completeness, we note that if we place the solutions we had already established in Chap. 9, that is $\hat{\delta}_1 = -0.973$ and $\hat{\delta}_8 = 0.973$, into Eqs. (11.11), (11.12), and (11.13), we obtain $16\hat{P}_{1.8|1} = 14$ and $16\hat{P}_{8.1|1} = 2$.

Estimating Item Parameters Simultaneously by Maximum Likelihood in the Presence of Missing Responses

With the notation and development above, we now generalize the procedure to the case of the design in Fig. 11.1 with three items in which only one item is common to both groups. As indicated above because not all persons have responded to all items, a design such as that one is often described as having *missing data* or *missing responses*.

The maximum likelihood estimation of the three items simultaneously requires the likelihood of all conditional responses. In the example, this is given by multiplying the conditional probabilities of the responses between items 1 and 8 and the responses between items 8 and 11. This gives

$$L = \prod_{n=1}^{16} \frac{e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{e^{-\delta_1} + e^{-\delta_8}} \prod_{n=17}^{23} \frac{e^{-x_{n8}\delta_8 - x_{n11}\delta_{11}}}{e^{-\delta_8} + e^{-\delta_{11}}} \tag{11.14}$$

where the product in the second term which has responses to items 8 and 11 is made to run from n = 17 to 23 because they are different persons from those who responded to items 1 and 8, which we have running from n = 1 to 16.

Then expanding Eq. (11.14)

$$L = \frac{e^{-\sum_{n=1}^{16} x_{n1}\delta_1 - \sum_{n=1}^{16} x_{n8}\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}} \frac{e^{-\sum_{n=17}^{23} x_{n8}\delta_8 - \sum_{n=17}^{23} x_{n11}\delta_{11}}}{(e^{-\delta_8} + e^{-\delta_{11}})^{7}} \tag{11.15}$$

and the log likelihood is

$$\begin{aligned} \ln L &= -\sum_{n=1}^{16} x_{n1}\delta_1 - \sum_{n=1}^{16} x_{n8}\delta_8 - \sum_{n=17}^{23} x_{n8}\delta_8 - \sum_{n=17}^{23} x_{n11}\delta_{11} - 16\ln(e^{-\delta_1} + e^{-\delta_8}) \\ &\quad - 7\ln(e^{-\delta_8} + e^{-\delta_{11}}) \\ &= -14\delta_1 - 2\delta_8 - 5\delta_8 - 2\delta_{11} - 16\ln(e^{-\delta_1} + e^{-\delta_8}) - 7\ln(e^{-\delta_8} + e^{-\delta_{11}}) \\ &= -14\delta_1 - 7\delta_8 - 2\delta_{11} - 16\ln(e^{-\delta_1} + e^{-\delta_8}) - 7\ln(e^{-\delta_8} + e^{-\delta_{11}}) \end{aligned} \tag{11.16}$$

It is evident that item 8, the item common to the two groups, is involved in more responses than the other two items which are responded to by only one group.

Using calculus, the equations that maximize the likelihood are

$$\delta_1 : -14 + 16\frac{e^{-\hat{\delta}_1}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = -14 + 16\hat{P}_{1.8|1} = 0 \tag{11.17}$$

$$\delta_8 : -7 + 16\frac{e^{-\hat{\delta}_8}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} + 7\frac{e^{-\hat{\delta}_8}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = -7 + 16\hat{P}_{8.1|1} + 7\hat{P}_{8.1|1} = 0 \tag{11.18}$$

$$\delta_{11} : -2 + 7\frac{e^{-\hat{\delta}_{11}}}{e^{-\hat{\delta}_8} + e^{-\hat{\delta}_{11}}} = -2 + 7\hat{P}_{11.8|1} = 0. \tag{11.19}$$

These equations are also not independent, and to obtain a particular solution, we impose the constraint

$$\hat{\delta}_1 + \hat{\delta}_8 + \hat{\delta}_{11} = 0. \tag{11.20}$$

Again, these equations are solved iteratively. We do not proceed to solve these equations in this way, but we leave it as an exercise to show that the solution in the last row of Table 11.2,

$\hat{\delta}_1 = -1.603$, $\hat{\delta}_8 = 0.343$, $\hat{\delta}_{11} = 1.259$, satisfies Eqs. (11.17), (11.18), (11.19) and (11.20).

The above method of maximum likelihood is called *conditional pairwise estimation*. It has some desirable properties, including that the estimates obtained converge to the correct estimates as the sample size increases. However, because the same item appears in different pairings, the responses are not totally independent, and therefore it is not used directly in tests of fit. We consider tests of fit in subsequent chapters.

The design of Fig. 11.1 generalizes so that, providing each item is paired with at least one other item in a data matrix, the estimation can be carried out. We refer to this point again in the last section of the chapter.

Equating Scores of Persons Who Have Answered Different Items from the Same Set of Items

We have considered, above, placing items on the same scale when not all persons have answered all items. Focusing now on persons, we recall that in the Rasch model all persons with the same total score will have the same proficiency estimate. This is because the total score is a sufficient statistic for the estimation of proficiency. However, if two persons have responded to different items, then because the difficulties of the items are different, persons with the same total score will have different proficiency estimates. Thus, if a person has attempted 20 relatively difficult items and has a score of 15, then that will give a greater proficiency estimate than if the person had attempted 20 relatively easy items and also has a score of 15.

To show how this appears in the estimation Eq. (10.5) from Chap. 10, it may be modified to

$$r_n = \sum_{i=1}^{I} a_{ni} x_{ni} = \sum_{i=1}^{I_n} a_{ni} \frac{e^{\hat{\beta}_n - \hat{\delta}_i}}{1 + e^{\hat{\beta}_n - \hat{\delta}_i}} \tag{11.21}$$

where I_n, with the subscript n, indicates the number of items person n has completed and a_{ni} is a dichotomous variable that takes on the value 1 if person n has responded to item i, and 0 if person n has not responded to item i. Thus, the sum on both sides of Eq. (11.21) only contains the items to which the person has responded.

Table 11.3 shows the items from the example in Chap. 3 analyzed as dichotomous items, as in Chap. 9. The items have been labelled and ordered in terms of their difficulties. The top part of Table 11.3 shows two subtests formed from two different sets of items, one with the easiest 9 items and one with the most difficult 9 items. The second part of the table shows the proficiency estimates on each of the possible scores from 1 to 8, with extrapolated values for 0 and 9. It is evident that for the same total score, the proficiency estimate on the more difficult items is greater than that from the easier items. Figure 11.2 shows the graphical relationship between the scale values of β and scores on the two tests.

In each case, the person's total score (on those items attempted) is the relevant statistic for estimating the proficiency, but the estimate itself depends on the difficulty (parameters) of the items. If the items are on the same scale, then the proficiency estimates will also be on the same scale.

Table 11.3 Person estimates from two sets of items on the same scale

Item subtests selections		
Item	Set 1	Set 2
2	X	
1	X	
5	X	
6.4	X	
3	X	
4	X	
6.3	X	
9.1	X	
9.3	X	
6.1		X
9.2		X
6.2		X
7		X
10.1		X
8		X
10.3		X
10.2		X
10.4		X
No.	9	9
Max	9	9
Total score and equivalent proficiencies		
Score	Set 1	Set 2
0	−3.642	−1.992
1	−2.764	−1.117
2	−2.088	−0.422
3	−1.572	0.122
4	−1.120	0.617
5	−0.688	1.112
6	−0.242	1.640
7	0.263	2.247
8	0.923	3.022
9	1.780	3.926

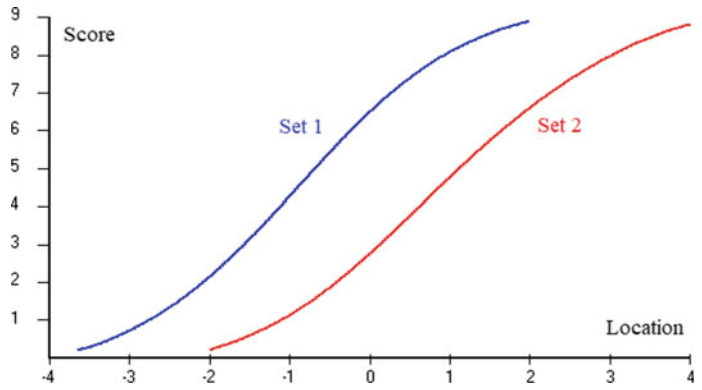

Fig. 11.2 Estimates of person proficiency from two tests composed of items on the same scale

Applications

Estimating item locations on the same scale where not all persons have responded to all items is common in education. For example, in large scale assessment exercises, at national and international levels, it may be necessary to compare the proficiencies of persons over different year groups, and older year groups need to be administered more difficult items than younger year groups. In order to link the items administered to the two groups, some common items, those which may be somewhat more difficult for the younger group but not too difficult, and somewhat easier for the older group but not too easy, may be administered as common items. Then all items are linked through the common items as shown above, and person estimates are obtained only from those items to which the persons have responded.

As a concrete example, in Australia, there is a National Assessment Program in Literacy and Numeracy (NAPLAN) in which students in years 3, 5, 7 and 9 are assessed and the assessments are placed on the same scale. This design of the assessments requires that there are some common items between adjacent year groups, with the majority of items unique to each year group.

Another example in education is where achievements over time are to be compared. It is important that the items from year to year are not the same. If they are, then performance on the items becomes an end itself and students and teachers can prepare just for those items. In this case, validity is destroyed, and improvements would be considered artificial. Instead, new items which assess the same variable need to be constructed. Then, the items from different times of assessment can be considered *illustrative* of achievement of the variable and the performance does not depend on which items have been chosen. To link the items over different times, it is necessary to have some items that are not made public and that are used across times. These items provide the link.

The above procedure for linking items is possible provided there is an overlap of persons and items so that there are no mutually exclusive blocks of persons and items.

The greater the overlap, the stronger the link. Once the link has been made and the item parameters have been estimated, then the person parameters can be estimated from the different subsets of items, and these estimates are on the same scale.

The above example of NAPLAN involves common items between adjacent year groups, with the older students being given more difficult items. If you recall reading Rasch's Chap. 1 in Rasch (1960), this is exactly the design he had in measuring students' progress in reading with older students being given more difficult texts to read, but with students mostly from adjacent year groups having some texts in common.

Having items on the same scale and having students answer only those items which are close to their own proficiencies, is the basis of computer adaptive testing. Here, students are administered items that are close to their proficiency and not those either too difficult or too easy. Styles and Andrich (1993) show an example in which items were administered in a computer adaptive testing format and two forms of a test were linked using the principles described above.

Most modern computer programs can cater automatically for data missing in the sense that not all persons have attempted all items. This means that, in principle, it is possible to equate the scores of two or more tests from a common set of items that have been compiled from the same joint analysis.

In CTT, the approach to equating is to take people from the same population, and preferably the same people, and administer them all the tests to be equated. The persons are then ordered by their total scores on the respective tests, and the cumulative percentages are calculated. Then scores on different tests which reflect the same cumulative percentage are taken to be equivalent. This procedure is referred to as equipercentile equating. Styles and Andrich (1993) compare a Rasch equating to an equipercentile equating from CTT. The advantage of using the Rasch model is that not all students need to be administered the same items.

In addition to examples in education, there are examples of linking items in the health outcomes areas. Here there have been many instruments constructed that attempt to assess the same health status and many have some similar or same items. In the cases where there are common items, it is possible to link these different instruments. Linking such instruments means that studies which have used the different instruments can be compared.

References

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by B. D. Wright (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Styles, I., & Andrich, D. (1993). Linking the standard and advanced forms of the Raven's Progressive Matrices in both the pencil-and-paper and computer-adaptive-testing formats. *Educational and Psychological Measurement, 53*(4), 905–925.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 57–60). Newbury Park, CA: Sage.