Chapter 9
Estimating Item Difficulty

The concepts of *standard error of an estimate* and *maximum likelihood estimate* are only briefly introduced here but elaborated in the next chapter. To consolidate the concept of sufficiency and its implication, introduced in previous chapters, this chapter shows the application of the total score in the estimation of the relative difficulties of two items with dichotomous responses. We show an application of Eq. (8.2) from Chap. 8 in the estimation of item difficulty.

Application of the Conditional Equation with Just Two Dichotomous Items and Many Persons

Estimating Relative Item Difficulties

We now show an elementary application of Eq. (8.2) from Chap. 8 in which the difference in difficulties between two dichotomous items is estimated. This equation is generalized in software when there are more than two dichotomous items and when the items are polytomous. We consider these generalizations in later chapters.

We use again the data in Table 3.1 of Chap. 3, but now focus on items 1 and 8, two dichotomous items with different facilities. We will estimate their relative difficulties. The responses for just these two items are reproduced in Table 9.1. However, now the persons have been reordered according to their total scores on these two items.

Recall from the previous two chapters, and from above, that the key response patterns in the case of two dichotomous items are those in which one item is correct and the other is incorrect, that is, where the total score on the two items is 1. Lines in Table 9.1 mark off the persons with a total score of 1.

There are 16 persons with a total score of 1, and of these, 14 have item 1 correct and item 8 incorrect, and 2 have item 8 correct and item 1 incorrect. The responses for these two items are rearranged in a two-way table in Table 9.2. They show the

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_9

Table 9.1 Responses of 50 persons to two items on a 10-item test

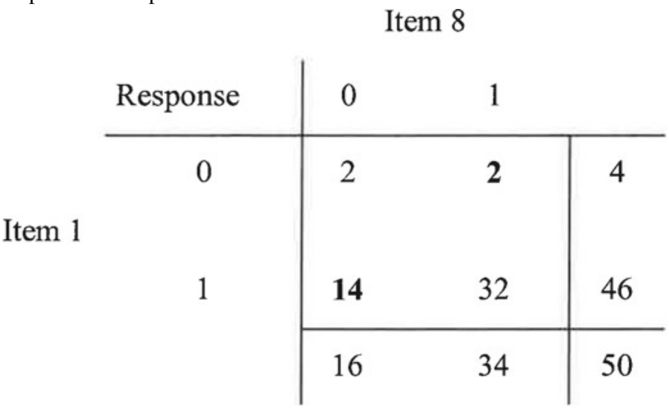
Person	1	8	Total score
2	0	0	0
38	0	0	0
8	1	0	1
10	1	0	1
11	0	1	1
13	1	0	1
17	1	0	1
23	1	0	1
25	1	0	1
27	1	0	1
29	1	0	1
30	0	1	1
35	1	0	1
41	1	0	1
42	1	0	1
43	1	0	1
44	1	0	1
45	1	0	1
1	1	1	2
3	1	1	2
4	1	1	2
5	1	1	2
6	1	1	2
7	1	1	2
9	1	1	2
12	1	1	2
14	1	1	2
15	1	1	2
16	1	1	2
18	1	1	2
19	1	1	2
20	1	1	2
21	1	1	2
22	1	1	2
24	1	1	2
26	1	1	2
28	1	1	2
31	1	1	2

(continued)

Table 9.1 (continued)

Person	1	8	Total score
32	1	1	2
33	1	1	2
34	1	1	2
36	1	1	2
37	1	1	2
39	1	1	2
40	1	1	2
46	1	1	2
47	1	1	2
48	1	1	2
49	1	1	2
50	1	1	2
Total:	**46**	**34**	
Facility:	**92**	**68**	
Discrimination:	**0.36**	**0.48**	

Table 9.2 Responses of 50 persons to items 1 and 8



		Item 8		
	Response	0	1	
Item 1	0	2	**2**	4
	1	**14**	32	46
		16	34	50

frequencies of all four patterns of responses for the 50 persons, with the responses with a total score of 1 in bold.

In order to estimate the relative difficulties of these two items using Eq. (8.2) from Chap. 8, we rearrange it and replace the subscript 2 for item 2 with the subscript 8 for item 8. Thus, the probability of item 1 correct and item 8 incorrect, given that the sum of the responses to the two items is 1, is

$$\Pr\{(x_{n1}=1, x_{n8}=0)|r_n=1\} = \frac{e^{-\delta_1}}{e^{-\delta_1}+e^{-\delta_8}} = \frac{e^{\delta_8-\delta_1}}{1+e^{\delta_8-\delta_1}} \tag{9.1}$$

We notice that this has the same structure as the dichotomous RM, except that the two parameters are the difficulties of the two items rather than an item parameter and a person parameter.

The probability of the complementary response, item 8 correct and item 1 incorrect, given that the sum of the responses to the two items is 1 is given by

$$\Pr\{(x_{n1}=0, x_{n8}=1)|r_n=1\} = \frac{e^{-\delta_8}}{e^{-\delta_1}+e^{-\delta_8}} = \frac{e^{\delta_1-\delta_8}}{1+e^{\delta_1-\delta_8}} = \frac{1}{1+e^{\delta_8-\delta_1}} \tag{9.2}$$

We have made the denominator $(1+e^{\delta_8-\delta_1})$ in Eq. (9.2) the same as that in Eq. (9.1). This means that in the ratio of Eqs. (9.1) and (9.2), this denominator will cancel. Thus, the ratio of Eqs. (9.1) and (9.2) is

$$\frac{\Pr\{(x_{n1}=1, x_{n8}=0)|r_n=1\}}{\Pr\{(x_{n1}=0, x_{n8}=1)|r_n=1\}} = \frac{(e^{\delta_8-\delta_1})/(1+e^{\delta_8-\delta_1})}{1/(1+e^{\delta_8-\delta_1})} = e^{\delta_8-\delta_1} \tag{9.3}$$

Taking the logarithm of both sides gives

$$\ln\left[\frac{\Pr\{(x_{n1}=1, x_{n8}=0)|r_n=1\}}{\Pr\{(x_{n1}=0, x_{n8}=1)|r_n=1\}}\right] = \delta_8-\delta_1 \tag{9.4}$$

This is the equation we use to estimate the difference $\delta_8-\delta_1$.

Before proceeding, we stress that the above equations, and in particular Eq. (9.4) which we use, do not have the person parameter β_n of any person. Thus, although the probabilities of a correct or incorrect response to both items depend on each person's parameter, Eq. (9.4) does not involve any person's parameter. This means that the 16 responses in Table 9.2 which are in bold are replications of each other in the sense that they are governed by the same parameter, in this case, the difference $\delta_8-\delta_1$.

Note Eq. (9.1) is a Bernoulli variable. This is because every response is either $\{(x_{n1}=1, x_{n8}=0)|r_n=1\}$ or $\{(x_{n1}=0, x_{n8}=1)|r_n=1\}$ with a complementary probability which sums to 1. We can formalize this observation by defining a new Bernoulli random variable a_{18} which takes the value $a_{18}=1$ when $\{(x_{n1}=1, x_{n8}=0)|r_n=1\}$ and $a_{18}=0$ when $\{(x_{n1}=0, x_{n8}=1)|r_n=1\}$. The subscript *18* in

a_{18} denotes reference to items 1 and 8. You may need to check *Statistics Review* 10 where random variables and Bernoulli random variables are defined.

Table 9.3 shows the responses from Table 9.1 for which $\{r_n = 1\}$ together with values for the random variable a_{18} and a count of the number of persons.

In Table 9.3, we have 16 Bernoulli replications with exactly the same probability of the response $a_{18} = 1$. This probability is given by Eq. (9.1) and is independent of any person's parameter which of course will all be different from each other. Even if two people obtain the same score, it does not mean that they have the same proficiency. They simply have the same score and we cannot distinguish between them. However, as we add more items, we increase our opportunity to distinguish between any two persons.

The sum of these Bernoulli variables gives a binomial variable. Therefore, we know that the estimate of the probabilities $\Pr\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\}$ is simply the mean of the number of responses $a_{18} = 1$, which is the proportion of responses $\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\}$.

The probability of the complementary response $\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\}$ is simply $1 - \Pr\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\}$ and its estimate is the complementary proportion of responses to that of $\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\}$. We may write

$$\text{Proportion}\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\} = \frac{14}{16}\text{and}$$

$$\text{Proportion}\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\} = \frac{2}{16}$$

Table 9.3 Responses of persons to two items given $\{r_n = 1\}$

Person	Person count	Item 1	Item 8	$\{r_n = 1\}$	a_{18}
8	1	1	0	1	1
10	2	1	0	1	1
11	3	0	1	1	0
13	4	1	0	1	1
17	5	1	0	1	1
23	6	1	0	1	1
25	7	1	0	1	1
27	8	1	0	1	1
29	9	1	0	1	1
30	10	0	1	1	0
35	11	1	0	1	1
41	12	1	0	1	1
42	13	1	0	1	1
43	14	1	0	1	1
44	15	1	0	1	1
45	16	1	0	1	1
	Total = 16				Sum = 14

Substituting these proportions as estimates of the respective probabilities in Eq. (9.4) gives

$$\ln\left[\frac{14/16}{2/16}\right] = \ln\left[\frac{14}{2}\right] = \ln[7] = \hat{\delta}_8 - \hat{\delta}_1.$$

That is,

$$\hat{\delta}_8 - \hat{\delta}_1 = \ln[7] = 1.946. \tag{9.5}$$

Thus item 8 is more difficult than item 1, and our estimate is that the difference is 1.946 logits. To designate that this is an estimate, the item parameters $\delta_8,\ \delta_1$ now have a 'hat' $\hat{\delta}_8,\ \hat{\delta}_1$.

One can make tangible the difference in difficulties of the two items by considering the proportion of persons who have item 1 correct, given that they have only one of items 1 and 8 correct. This proportion is 14/16, that is 0.875, which is relatively large. Complementary to this proportion, the proportion of persons who have item 8 correct given that they have only one of items 1 and 8 correct, is 2/16, that is 0.125. This is a small proportion. Clearly, item 1 is substantially easier than item 8.

An important point to notice, and to understand, is that this difference is not the same as if we only considered the number of persons who answered each item correctly. It is evident from Table 9.2 that 46 persons answered item 1 correctly and 34 answered item 8 correctly. These are respective proportions of 0.92 and 0.68. Thus although item 8 shows itself to be more difficult than item 1, as in the above calculations, it appears to be closer in difficulty if only the number correct is considered. The reason for this can be explained by considering the high number of very proficient persons who answered both items correctly. When these are included in the overall calculation of difficulty (or facility), the difference in difficulties of the items appears smaller than in the conditional estimation given the total score. Thus, suppose that there were another 20 persons in the sample who were very proficient and that they answered both items correctly. Then the numbers correct would be respectively 66 and 54 from a total of 70 persons. The proportions correct are respectively 0.94 and 0.77, suggesting an even smaller difference in difficulties between the two items.

In these latter calculations, the apparent relative difficulties are affected by the proficiencies of the person; while the calculation conditional on the total score of 1 is not affected by these proficiencies.

Estimating Person Proficiencies

We have stressed that in the dichotomous RM, the sufficiency of the total score for the person's proficiency implies that all the information regarding this proficiency is in the total score, and no further information is in the response pattern. We consider the estimation of the person proficiency in the next chapter.

An Arbitrary Origin and an Arbitrary Unit

The Arbitrary Origin

We noted it incidentally above, but it is essential to appreciate that we have estimated only the *difference* between items 1 and 8. We cannot give each item its own independent difficulty estimate. However, for purposes of efficiency, we can give each its own value by setting an *arbitrary origin*.

In any analysis, this is generally done by setting the sum of the item parameters to zero.

In the above example, we set

$$\hat{\delta}_8 + \hat{\delta}_1 = 0 \tag{9.6}$$

Then by adding Eqs. (9.5)–(9.6), we have

$$2\hat{\delta}_8 = 1.946; \quad \hat{\delta}_8 = 0.973,$$

and by subtracting Eq. (9.5) from Eq. (9.6), we have

$$2\hat{\delta}_1 = -1.946; \quad \hat{\delta}_1 = -0.973.$$

Now we can write that $\hat{\delta}_1 = -0.973$ and $\hat{\delta}_8 = 0.973$ recognizing that this origin of 0, to which each value is referenced, is indeed arbitrary.

Although the origin in any analysis is arbitrary and is generally set to 0, it is often convenient to set it to some other value. For example, if a test has been defined in some previous application, and new items are added to the test, then the new items need to be referenced to the same origin as the previous application. This can be done in a number of ways, with only one constraint the equivalent of Eq. (9.6) required. For example, suppose a test composed of some items from a previous administration and some new items is administered to a group of people. Then, the analysis can be performed with the mean difficulty of the previous items fixed to their difficulty on the previous administration. Fixing this mean retains the origin of the previous administration and the difficulty estimates of the new items will have the same origin as the previous administration.

The choice of origin affects the proficiency values of the persons that are estimated with the items. For example, suppose the arbitrary origin of 0 was changed, to avoid negative numbers, to be say 50. That implies that 50 was added to the estimated value of each item. Because the difference $\beta_n - \delta_i$ must remain constant, each person's proficiency must also have the value 50 added to it. For example, for a group of persons, the mean would be adjusted from whatever its value, $\bar{\beta}$, might be from an analysis in which the origin of the items is 0, to have 50 added to it.

The Arbitrary Unit

The arbitrary origin is more visible than the arbitrary unit. This is because each analysis has to make this explicit. To see the role of the arbitrary unit, consider the original equation, Eq. (6.5), of the dichotomous Rasch model from Chap. 6 which is reproduced below

$$\Pr\{x_{ni} = 1|\beta_n, \delta_i\} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \tag{9.7}$$

Equation (9.7) can be written as

$$\begin{aligned}\Pr\{x_{ni} = 1|\beta_n, \delta_i\} &= \frac{e^{\alpha(\beta_n/\alpha - \delta_i/\alpha)}}{1 + e^{\alpha(\beta_n/\alpha - \delta_i/\alpha)}} \\ &= \frac{e^{\alpha(\beta_n^* - \delta_i^*)}}{1 + e^{\alpha(\beta_n^* - \delta_i^*)}}\end{aligned} \tag{9.8}$$

where $\alpha > 0$ is an arbitrary real number,

$$\beta_n^* = \beta_n/\alpha; \quad \delta_i^* = \delta_i/\alpha \tag{9.9}$$

without changing the value of the probability in Eq. (9.7). In general, we leave the value $\alpha = 1$ in writing the equation and in estimation. That is why it is not as conspicuous as the specification of the constraint that provides the origin. The value of α needs to be greater than 0, otherwise, the item does not operate in the same way as the other items. However, that it can be given different values is the sense in which the unit is arbitrary. It is stressed that it is the *expression* of the values of the person and item parameters that is arbitrary.

For example, if we specify that $\alpha = 2$ in the analysis of the data set of Table 9.3, then the estimates would be given by

$$\begin{aligned}2/(\hat{\delta}_8/2 - \hat{\delta}_1/2) &= 2(\hat{\delta}_8^* - \hat{\delta}_1^*) \\ &= \ln[7] = 1.946\end{aligned} \tag{9.10}$$

from which

$$(\hat{\delta}_8^* - \hat{\delta}_1^*) = 1.946/2 = 0.973 \tag{9.11}$$

If each item were to be given a single value by imposing the arbitrary origin $\hat{\delta}_8^* + \hat{\delta}_1^* = 0$, then

$$\hat{\delta}_8^* = 0.4865 \text{ and } \hat{\delta}_1^* = -0.4865.$$

As with the origin, which sometimes needs to be defined given some previous administration of a test, the unit might also have to be defined given a previous administration of a test with new items. As indicated above, one constraint such as the mean of the item difficulties of some set of items in a joint analysis with new items can retain the origin. To retain the unit from a previous administration of some items, fixing the difficulties of just two items is sufficient. However, if there are more than two previous or original items in a data set with some new items, as is usually the case, then all their values might be fixed to that from the previous analysis. Another option is to fix, not only their mean, but their standard deviation from a previous analysis. Then the estimates of the remaining items are in the same unit as the original items. In summary, to fix the origin one constraint is needed on the sum of the difficulties, and to fix the unit one constraint is required on the spread of the difficulties.

Generalizing to Many Items

As indicated above the equations for estimating the item parameters, conditioning out the person parameters can be generalized for purposes of estimating the responses of many persons to many items. There are two ways of proceeding. One is to proceed by considering all possible combinations of pairs and build up an equation that way. That is, the pairwise structure in Table 9.2 is built up with all the pairs of items. The other one is to extend Eqs. (8.3) and (8.4) from Chap. 8. The former is easier in some ways but has some disadvantages, and the latter which follows, more rigorous theoretically, is more complicated and it too has some disadvantages. However, from an estimation point of view only, for the same items, as the sample size of persons increases both converge to the same estimates. In practice, when data approximate the model, they are also virtually indistinguishable.

Maximum Likelihood Estimate (MLE)

Equation (9.4) which we used to estimate the difference between the difficulties of items 1 and 8 needs to be generalized in a different way when there are more items. Equation (9.4) is referred to as a *maximum likelihood estimate*. It is the complementary feature of sufficiency formulated by Fisher. Although MLE is different from the least squares estimate, we considered for fitting a regression equation in *Statistics Review* 4, it has the same idea. In the case of a least squares estimate, the estimated values of the parameters of the linear model are such that the sum of squares of the deviations about the linear regression line are a *minimum*. In the MLE, the estimated value of the parameter is the one which maximizes the probability that this set of responses is observed according to the model. This probability of a set of responses is called a *likelihood*. Because it requires calculus to find the maximum value of a function, we do not derive this equation here. We show a little more of the MLE in the next chapter when we consider the estimation of the person locations, given the item location estimates are taken as known.

Because the equations first involved *conditioning* on the total score, and eliminating the person parameter, the estimation is known as *conditional* maximum likelihood estimation.

Item Difficulty Estimates

The difficulties of all of the items of the example in Chap. 3 taken as dichotomous items are displayed in Table 9.4. The method of estimation is based on the first of the above generalizations from the estimation of the relative difficulties of just two items. That is, a table is formed for a pair of items just like Table 9.2, for example in this case items 1 and 8. Then taking item 1 as the focus first, a table such as Table 9.2 is made up for item 1 in relation to every other item. The statistic for item 1 then is the sum, over all item pairs, of the number of times this item has a response of 1 when the response to the other item is 0. Then such a table is formed for every item in relation to the other items.

Table 9.4 Difficulty estimates for dichotomous items

Item	Linear total	Location	SE	Total score
2	209	−2.004	0.715	48
1	194	−1.296	0.544	46
5	195	−1.264	0.538	46
6.4	204	−1.100	0.508	46
3	195	−0.672	0.443	44
4	170	−0.676	0.443	43
6.3	182	−0.539	0.426	43
9.1	179	−0.335	0.403	42
9.3	175	−0.338	0.403	42
6.1	166	−0.220	0.391	41
9.2	153	0.063	0.366	39
6.2	137	0.435	0.341	36
7	144	0.421	0.342	36
10.1	133	0.551	0.335	35
8	124	0.598	0.333	34
10.3	70	1.573	0.314	24
10.2	47	2.269	0.330	16
10.4	37	2.535	0.343	15

Table 9.4 shows for each item a *Linear Total* and a *Total Score*. The former is the total we referred to above, the number of times an item has a positive response when another item has a negative response, summed over all the pairs of items. The latter is simply the number of positive (correct) responses for each item, which we had calculated as part of the Guttman analysis of the responses. The items are ordered according to their *total score*, not the linear total. These two totals are not identical, but their order is very close. Because the estimation uses the *linear total*, not the *total score*, the *total scores* of the items are not in exact correspondence with their relative difficulties. However, again they are close. The second method we mentioned above, that which generalizes to many items, does give difficulty estimates which are exactly in the same order as the total scores. However, if the responses fit the Rasch model, then as the sample size increases, the two kinds of estimates get closer and closer together, they converge. In practice, as indicated above, they are very close to each other and well within the standard error of the estimate of the item difficulty.

The *standard errors* of the estimates of the item locations (difficulties) are also shown in Table 9.4. They also arise directly out of maximum likelihood estimation theory. We revisit them in the next chapter when we consider the estimation of the person parameters, given that we have used the conditional method of estimating the item parameters while eliminating all the person parameters. Sometimes the process of estimating the item difficulties, which locates the items on the continuum, is called test or item *calibration*. Then the process of estimating the person proficiencies, which locates the persons on the same continuum, is termed *person measurement*.

Exercises

1. Estimate the relative difficulties of item 1 and item 2 from the data set used in the Exercises at the end of Chap. 3 using the process shown above.
2. What are the estimates if the difficulties need to be expressed in ½ of the unit that appears when $\alpha = 1$.
3. Suppose that both the origin and the unit need to be specified to an a priori value. Specifically, suppose that the mean of the item difficulties needs to be 10 and that the unit, as in 2 above, is ½ of the unit that appears when $\alpha = 1$. What are the difficulty estimates?

Further Reading

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

Humphry, S., & Andrich, D. (2008). Understanding the unit implicit in the Rasch model. *Journal of Applied Measurement, 9,* 249–264.