Chapter 13
Fit of Responses to the Model I—Item Characteristic Curve and Chi-Square Tests of Fit

Statistics Review 11: The Chi-square test

There are typically two aspects to fit of responses to the model that need to be considered: how the items fit the model and how the persons fit the model. Fit can be assessed graphically and also formally through the use of statistics. This chapter involves two parts: (1) a review of the Item Characteristic Curve (ICC) as a graphical test of item fit including comparing observed proportions in class intervals with the ICC; (2) the χ^2 test as a statistical test of fit between the data and the ICC. The fit-residual statistic to assess both person and item fit will be discussed in a subsequent chapter.

Statistics Review 13: Distribution theory

Part II of this book consists of more advanced concepts in Rasch measurement. A number of chapters deal with violations of the Rasch model and how these are revealed in tests of fit. Distribution theory is very important in understanding and carrying out tests of fit. In order to understand tests of fit review distribution theory in *Statistics Review 13*.

A Graphical Test of Item Fit

The Item Characteristic Curve (ICC)

In Classical Test Theory (CTT) the empirical check that the items are working as expected is carried out by calculating the discrimination index. This was discussed in Chap. 3. In CTT there is no special criterion as to what is a good discrimination and what is a bad one—it is simply the case that the greater the discrimination the better. You will see that with Rasch Measurement Theory (RMT), this idea can be refined substantially. We continue with items or tasks that are scored dichotomously

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_13

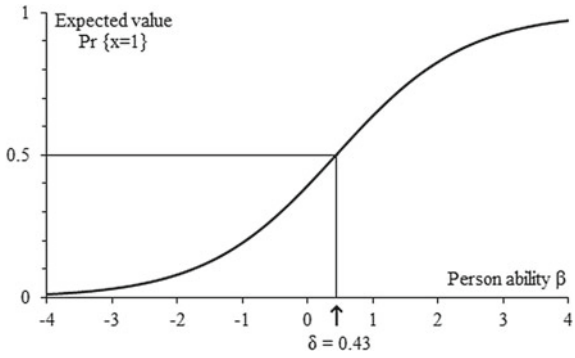

Fig. 13.1 Item characteristic curve for dichotomous item 6.2 of Table 5.3

(right or wrong), and then in Chap. 20 we will see how the same ideas can be applied to ratings of performance where partial credit is given.

You were introduced to the item characteristic curve (ICC) in Chap. 6. This is the probability that a person n with given proficiency β_n responds correctly to an item i with difficulty δ_i. Such a curve is reproduced in Fig. 13.1 for item 6.2 from the data in Table 5.3 in Chap. 5. Notice that where the proficiency $\beta_n = \delta_{6.2} = 0.43$, the probability of a correct response is equal to 0.5.

The curve in Fig. 13.1 is a theoretical curve for a given item difficulty. It shows the probability that a person with any particular proficiency will answer the item correctly. We also know that this probability is a theoretical proportion of the number of correct responses. It is also the theoretical average of the number of persons who answered the item correctly. If we had many people with the same total score, and therefore the same proficiency estimate, we could compare the observed proportion correct on the item with the theoretical probability. Often we do not have enough people with the same total score for the entire score range. We can, however, form class intervals exactly as we did in analyzing data according to the Guttman structure.

We now proceed to elaborate the Guttman analysis in terms of the Rasch model for the data from Table 5.3 of Chap. 5, as below.

Observed Proportions in Class Intervals

From the proficiency of each person, we form class intervals and calculate the average proficiency. This is similar to finding the average of the raw scores in order to locate each class interval, but rather than the raw scores it is the estimated proficiencies that are used. Now, we simply call them *class intervals* because sometimes we may wish to have more than three class intervals.

We continue with the example listed in Table 10.3 of Chap. 10. However, for the person estimates, and to illustrate the effect of the bias mentioned in Chap. 10, the person estimates are what is referred to as weighted likelihood estimates. The

RUMM2030 Interpreting Manual (Andrich, Sheridan, & Luo, 2018) describes these estimates in more detail. In Table 10.3, the persons are ordered according to their proficiency and items are ordered according to their difficulty. Because of a special feature in the way the difficulties of the items are estimated, and to ensure there is no bias in the estimates, it turns out that the items are not in the exact same order as are the items in the original Guttman analysis in Table 5.3 of Chap. 5. Items 3 and 4 which are very close to each other in difficulty change their order, as well as items 6.2 and 7. The effect is relatively small, and only when items are very close to each other can the difficulties appear in an order not quite the same as the order of their total scores. It is also a sign that the data do not fit the model perfectly. If the sample was very large, and the data fitted the model perfectly, then this reversal would not occur even for items that are close together in difficulty. In that case, if persons answer the same items, then they will always be ordered according to their total scores.

For convenience, and for comparison with the previous analysis, we again form three class intervals as in the Guttman analysis. Table 13.1 reproduces Table 10.3, but instead of the standard error for each person, three class intervals are formed.

In this case, the class intervals have the same persons as in the Guttman analysis in Table 5.3 of Chap. 5. The program RUMM2030 places people into class intervals in such a way that they are as close to equal in size as possible. A difference between Table 13.1 and the Guttman analysis of Chap. 5 is that the person who answered all the items correctly is not included in this analysis.

The proportions of people who answered each item correctly in each class interval are calculated as in the Guttman analysis. These are shown for item 6.2 in Table 13.2. However, in addition to the observed proportion of people who answered the item correctly in each class interval, we now have an expected proportion correct according to the Rasch model—this is the estimated probability shown in Table 13.2.

Figure 13.1 is now repeated in Fig. 13.2, but the proficiencies for each of the class intervals and the proportion of persons who answered the item correctly in that class interval are also shown.

The essential difference between Fig. 13.2 and the Guttman analysis is that in Fig. 13.2 we have a theoretical curve against which to compare the proportions of persons in each class interval who answered the item correctly. In the Guttman analysis, we did not have such a curve. All we knew was that we would want these proportions to increase as the total scores of the persons in the class intervals (that is, their average proficiencies) increased.

Item 6.2 is an item whose discrimination is excellent—even a bit "too good". The proportions are a little steeper than the theoretical curve. We come back to this point later in this chapter and again in the next chapter. Figures 13.3 and 13.4 show similar information for items 9.2 and 9.3. Item 9.3 does not discriminate very well—the observed proportions are flatter than the theoretical curve. However, this is in part because the item is very easy and all the class intervals have a high mean.

It is important that you appreciate the two ways in which these figures differ from the Guttman structure.

(1) Unlike the Guttman analysis, there is a theoretical curve as a criterion.
(2) Unlike the Guttman analysis, where the raw scores are averaged in the class intervals, the proficiencies are estimated first and then the average is taken.

In formulating a model for data, it is expected that the data will accord well with the model. Recall that the Rasch model is a theoretical model based on the requirement of invariant comparisons. However, when the data do not accord with the model, then the model can still be very useful in understanding the data. It helps to diagnose where the data are different from what was expected from the model. Usually, there is an explanation for such effects. Often, experience can tell you what has gone wrong quickly. However, equally often one needs to know the test, the

Table 13.1 Table 10.3 formed into three class intervals

Person	Responses	Total score r_n	Location $\hat{\beta}$ (WLE)	Class interval average proficiency $\bar{\hat{\beta}}$
38	1011010010100000000	6	−0.889	0.311
2	1011011101000000000	7	−0.608	
40	0101111100000101000	8	−0.335	
42	1100111111100100000	10	0.209	
41	1111011111010000000	10	0.209	
44	1111011111110000000	11	0.493	
8	1111101101011100000	11	0.493	
35	1111110111011000000	11	0.493	
11	1011111111100110000	12	0.795	
9	1101111110110110000	12	0.795	
46	1110110110110110100	12	0.795	
29	1111011110111100000	12	0.795	
25	1111101011111100000	12	0.795	
27	0111011111011001111	13	1.123	1.289
18	1101111101111010001	13	1.123	
36	1110111101111110000	13	1.123	
37	1111011111111010000	13	1.123	
20	1111100111111011000	13	1.123	
48	1111101011111110000	13	1.123	
13	1111110110111101000	13	1.123	
34	1111110111001111000	13	1.123	
32	1111111010111110000	13	1.123	
22	1111111011011010001	13	1.123	

(continued)

Table 13.1 (continued)

Person	Responses	Total score r_n	Location $\hat{\beta}$ (WLE)	Class interval average proficiency $\bar{\hat{\beta}}$
43	111111111101110000	13	1.123	
14	111110101111011110	14	1.492	
12	111111100101011111	14	1.492	
15	111111100110111110	14	1.492	
21	111111111000111110	14	1.492	
5	111111110010011110	14	1.492	
4	111111111110011100	14	1.492	
16	111111111110111000	14	1.492	
45	111111111111000011	14	1.492	
17	111111111111100100	14	1.492	
7	111111110111111100	15	1.920	2.470
50	111111111110011110	15	1.920	
49	111111111110111100	15	1.920	
23	111111111111110001	15	1.920	
6	111111111111111000	15	1.920	
24	111111111111111000	15	1.920	
33	111110111111111101	16	2.445	
26	111111011111111101	16	2.445	
10	111111111110110111	16	2.445	
31	111111111111101110	16	2.445	
19	111111111111111010	16	2.445	
30	101111111111111111	17	3.153	
28	111011111111111111	17	3.153	
1	111111111111101111	17	3.153	
39	111111111111111101	17	3.153	
47	111111111111111101	17	3.153	
3	111111111111111111	18	$+\infty$	

Table 13.2 Proportion of correct responses for item 6.2 in each class interval

Item	Proficiency	Observed proportion correct	Estimated probability correct
CI_1	0.311	0.38	0.47
CI_2	1.289	0.75	0.70
CI_3	2.470	0.94	0.87

population, and the test conditions in order to understand any discrepancies between the model and the data.

When we have a theoretical curve, it is evident that when the observed proportions deviate substantially from this theoretical curve, then we have some kind of misfit between the data and the model. It is relevant to appreciate that the discrimination of the theoretical curve is the average discrimination of all the items. This provides the frame of reference to study an item with greater or smaller discrimination, and then substantial deviations are seen as *outliers*.

There are three kinds of ways that the observed proportions might deviate from the theoretical values, which are given as follows:

1. The observed proportions are *flatter* than the theoretical curve, in which case the item does not discriminate enough. Item 9.3 is such an item.
2. The observed proportions are *haphazardly and substantially different* from the theoretical curve. This requires specific interpretation with knowledge of the construct.

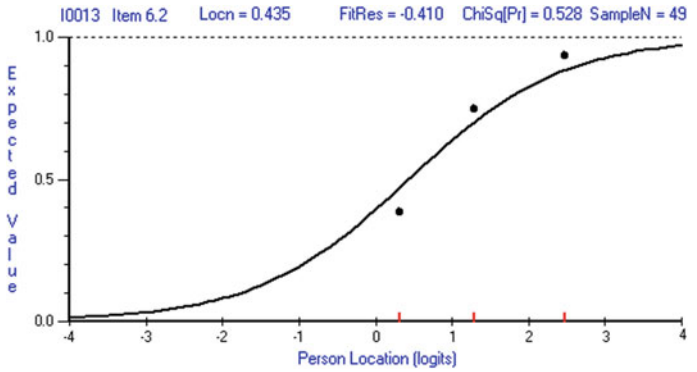

Fig. 13.2 ICC and proportions correct in three class intervals for item 6.2

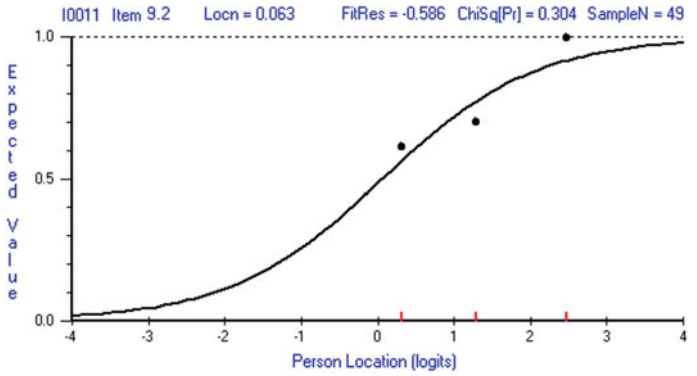

Fig. 13.3 ICC and proportions correct in three class intervals for item 9.2

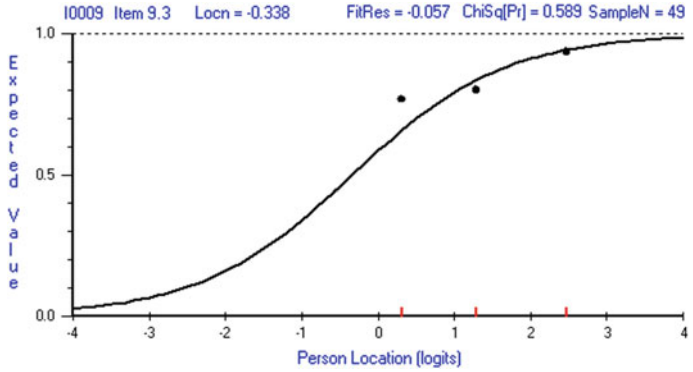

Fig. 13.4 ICC and proportions correct in three class intervals for item 9.3

3. The observed proportions are *steeper* than the theoretical curve. This means that the discrimination is greater than expected. Item 6.2 in Fig. 13.2 is an example of such an item. This is another difference between CTT and Rasch measurement theory (RMT). In the former, the greater the discrimination the better. In the latter, when the observed proportions are systematically greater than the theoretical proportions, then we also show concern. Rasch himself was concerned indirectly with this case; in principle, it shows that there is a greater dependence among responses in one form or another.

1 and 3 describe cases of *systematic* misfit and 2 describes *non-systematic* misfit.

A Formalised Test of Item Fit—χ^2

To check if the data do accord with the model, we compare the expected number of correct responses in each class interval according to the model with the actual observed number of correct responses. This kind of comparison is called a *test of fit*. We can now formalize a test of fit between the data and the model for each item as a statistical χ^2 test as follows:

(1) We find the number of people who answered item i correctly in class interval g. This number is called T_{gi}.

(2) We find the number of people who are expected to answer the item correctly by first finding the probability that the persons in each class interval answered each item correctly. This number is given by the probability that the people in each class interval would answer the item correctly times the number of people in the class interval. Recall that the probability is simply a theoretical proportion. For example, if the probability is 0.3, and there were 10 people in the class interval, then the number who should have answered it correctly would be $(10)(0.3) = 3$.

In general, if N_g is the number of people in each class interval g, and P_{gi} is the theoretical proportion, that is the probability, that a person in class interval g answers item i correctly, then $N_g P_{gi}$ is the expected number of people in that class interval who answer item i correctly.

(3) For any item i, the difference between the observed number who answered the item correctly and the expected number is formed.
This may be written as

$$T_{gi} - N_g P_{gi} \tag{13.1}$$

where N_g is the number of persons in class interval g, P_{gi} is the probability that a person in class interval g will answer item i correctly, and T_{gi} is the number of persons in class interval g who do answer item i correctly.

In the graphical analysis, we compare the observed proportion with the estimated probability. For the formal statistical analysis, it turns out to be convenient to consider the total number correct rather than the proportion—however, the interpretation is the same.

Second, this difference is divided by the standard deviation of the number who are likely to answer the item correctly. This gives more tangible meaning to the difference and gives a standardized residual Z_{gi}. It is a standard score and may be expressed as

$$Z_{gi} = \frac{T_{gi} - N_g P_{gi}}{\sigma_{gi}} \tag{13.2}$$

where $\sigma_{gi} = \sqrt{N_g P_{gi}(1 - P_{gi})}$ is the standard deviation of the number correct.

The greater the standardized difference, the less likely the item will fit the model.

One could compare each of the standardized residuals against a standardized normal deviated from the normal distribution, and if it were greater than about $+2$ or less than -2 we could show concern, and that is in fact carried out.

However, to obtain an index for an item as a whole, these residuals are simply squared and added up and this gives an approximate χ^2 distribution on G $-$ 1 class intervals where G is the number of class intervals, i.e.

$$\chi_i^2 = \sum_{g=1}^{G} z_{gi}^2 \tag{13.3}$$

This number can be compared to the values of a theoretical χ^2 distribution on the specified degrees of freedom. This comparison can tell how likely it is that a χ^2 of this value or greater is to occur by chance.

You could make these comparisons by looking up a table. However, all of this information is provided in RUMM2030. Below is an interpretation of the RUMM2030 χ^2 test of fit output.

Interpretation of Computer Printout—Test of Fit Output

Below is a printout of the information that is provided by RUMM2030 for Item 6.2. This is followed by an explanation of each of the symbols in the table.

Item 6.2 (I0013) Locn = 0.435								
GROUP		LOCATION		COMPONENT		Category Responses		
No	Size	Max	Mean	Residual	ChiSqu		0	1
1	13	0.795	0.311	−0.670	0.448	OBS.P	0.62	0.38
						EST.P	0.53	0.47
OM = 0.38 EV = 0.47 OM-EV = −0.09 ES = −0.19						OBS.T		0.38
2	20	1.492	1.289	0.489	0.239	OBS.P	0.25	0.75
						EST.P	0.30	0.70
OM = 0.75 EV = 0.70 OM-EV = 0.05 ES = 0.11						OBS.T		0.75
3	16	3.153	2.470	0.769	0.592	OBS.P	0.06	0.94
						EST.P	0.12	0.88
OM = 0.94 EV = 0.87 OM-EV = 0.06 ES = 0.19						OBS.T		0.94
AVE = 0.71								
ITEM: df = 2 ChiSqu = 1.279 Significance = 0.528								

Note The value of EV and EST.P for Category Response 1 (in the dichotomous model) might not be identical due to rounding errors

Item 6.2 (I0013): **13** is the order of the item, and **Item 6.2** is the label we have given to this item.

Locn = 0.435: **Locn** is short for **location** of the item on the continuum, and it is the same as the difficulty of the item.

GROUP: This is the class interval. **No** is the group number or class interval. **Size** is the number of people in the group. Group 1 has 13 people in it, group 2 has 20 and group 3 has 16.

LOCATION: This is the person variable. It could have been called the person location. **Max** is the maximum proficiency of the group. It can help to check where the cut-off for the interval has been made by the computer program. This is 0.795 for group 1 (class interval 1). **Mean** is obviously the group's average location or proficiency, which is 0.311 for group 1.

COMPONENT: This refers to the components of the Chi-square statistic. **Residual** is the standardized difference between the observed number of persons in the group who have answered the item correctly and the expected number according to the model. This has a value of −0.670 for group 1. The equation for this value is given in *Statistics Review 11*.

Chi Squ: This is the Chi-square component for the group or class interval. It is simply the square of the residual value. This has a value of 0.448 for group 1.

Category Response: This indicates the response category. 0, 1 indicate that the possible scores for the item are 0 and 1. You will see when we deal with partial credit or rated items that these numbers can extend to 0, 1, 2 and so on.

OM: This is the observed mean for the class interval, expressed as a proportion. This value is 0.38 for group 1.

EV: This is the expected mean according to the model. This value is 0.47 for class interval 1.

OBS.P: This is the proportion of persons who responded with the scores of 0 or 1 in the class interval. This value is 0.62 for the score of 0 and 0.38 for the score of 1 for group 1. Note that with dichotomous responses, where scores can be only 0 or 1, that these proportions sum to 1; that is $0.62 + 0.38 = 1.00$. In the dichotomous case, the value for the score of 1 is the same as the OM.

EST.P: This is the estimated probability of persons who responded with the scores of 0 or 1 in the class interval. This value is 0.53 for the score of 0 and 0.47 for the score of 1 for group 1. Note again that with dichotomous responses, where scores can be only 0 or 1, that the sum of these probabilities also adds to 1.0; that is $0.53 + 0.47 = 1.00$. In the dichotomous case, the value for the score of 1 is also the same as the EV.

OBS.T: This is the probability of the response of 1 given that the response is either 0 or 1. In the case of the dichotomous item, this is the same as the probability of the response of 1, but it is different in the case of items with more than two categories.

OM-EV: This is the difference between the observed mean and the expected value.

ES: (Optional) This is a special standardized difference between OM and EV which does not take into account the size of the class interval.

df: This is the degrees of freedom for the Chi-square test. In this case, where there are three class intervals, the number of degrees of freedom is $3 - 1 = 2$.

Chi Squ: This is the total Chi-square for the item. For item 6.2 it is $0.448 + 0.239 + 0.592 = 1.279$.

Significance: This indicates the probability that a value as large as this would occur by chance if the responses fitted the model. In this case, it is evident that the probability is very high (0.528) that this value could have occurred by chance. That means that this item does fit the model very well. If the value were less than 0.01, then it would be considered unlikely to fit the model. This statistic, however, needs to be interpreted with some experience. It only approximates a Chi-square statistic and is inflated when the estimated probabilities are close to 0 or 1, and increases with the sample size. It is better to use it as an order statistic to see which items show much larger values than others, and to look at the graph such as the one in Fig. 13.2. It is also affected by how the groups are formed, although with large groups this should not have a large effect. In this item, the observed proportions are close to the theoretical curve.

The χ^2 statistic calculated as shown above is an excellent approximation for its purpose. However, it is sensitive to sample size, and therefore the same magnitude of discrepancies between the observed and expected frequencies will show as significant with increasing sample size. Here the graphical evidence should be taken into

account. In addition, the perspective that the ICC reflects the average discrimination can be exploited. The items can be ordered by the magnitude of their χ^2 values and those with large values can be seen as outliers. Sometimes just one or two items stand out as outliers. In such cases, the content and format of the items needs to be considered in interpreting the outliers. Sometimes the source of its misfit is an incorrect key for the correct answer in a multiple-choice item.

Exercises

Exercise 2: Basic analysis of dichotomous and polytomous responses in Appendix C.
Exercise 3: Advanced analysis of dichotomous responses Part A in Appendix C.

Reference

Andrich, D., Sheridan, B. E., & Luo, G. (2018). *RUMM2030: Rasch unidimensional models for measurement. Interpreting RUMM2030 Part III Estimation and statistical techniques*. Perth, Western Australia: RUMM Laboratory.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 63–67). Newbury Park, CA: Sage.
Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15–29.