Chapter 15
Fit of Responses to the Model II—Analysis of Residuals and General Principles

The first part of this chapter concerns an analysis of residuals where the focus is on the response of each person to each item. In particular, given the parameter estimates, the residual is formed between the response of a person to an item and the expected value according to the model. The estimates are compared and a new residual is constructed to summarize the fit of an item, or a person's profile, to the model. The second part of this chapter contains some general principles for assessing the fit of responses to the model.

The Fit-Residual

The residual of the response x_{ni} of each person n for each item i is simply

$$x_{ni} - E[x_{ni}], \tag{15.1}$$

wherein the case of the dichotomous Rasch model

$$E[x_{ni}] = \Pr\{x_{ni} = 1\} = P_{ni}. \tag{15.2}$$

The residual itself is a difference. To assess whether the magnitude is large or not, it is referenced to its standard deviation, $\sqrt{V[x_{ni}]}$. Therefore, the *standardized residual*

$$z_{ni} = \frac{x_{ni} - E[x_{ni}]}{\sqrt{V[x_{ni}]}} \tag{15.3}$$

is formed where $V[x_{ni}] = E[x_{ni}^2] - (E[x_{ni}])^2$ is the variance of x_{ni}.

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_15

The theoretical mean over an imagined infinite number of replications is $E[z_{ni}] = 0$ and $V[z_{ni}] = 1$. Because of the estimation equations for the person parameters, the sum of the standardized residuals will always be close to 0. In maximum likelihood estimation, it is exactly 0.

Therefore, to assess the magnitude of the residuals, these are squared to give z_{ni}^2. From these squared residuals, we obtain a summary value for a person and a summary value for an item by summing over the items *or* persons, respectively:

$$y_n^2 = \sum_{i=1}^{I} z_{ni}^2, \tag{15.4}$$

$$y_i^2 = \sum_{n=1}^{N} z_{ni}^2. \tag{15.5}$$

These summary values now need to be compared to their expected values—their expected values are their degrees of freedom. Therefore, a single person *or* item residual value that summarizes all the person–item residuals are respectively given by

$$y_n^2 - E[y_n^2] = \sum_{i=1}^{I} z_{ni}^2 - \sum_{i=1}^{I} f_{ni}, \tag{15.6}$$

$$y_i^2 - E[y_i^2] = \sum_{n=1}^{N} z_{ni}^2 - \sum_{n=1}^{N} f_{ni}. \tag{15.7}$$

These residuals can be standardized by dividing by their respective standard deviations:

$$Z_n = \frac{y_n^2 - E[y_n^2]}{\sqrt{V[y_n^2]}}, \tag{15.8}$$

$$Z_i = \frac{y_i^2 - E[y_i^2]}{\sqrt{V[y_i^2]}}. \tag{15.9}$$

Approximations for the Degrees of Freedom

In the RUMM2030 Interpreting Manual (Andrich, Sheridan, & Luo, 2018) there is a discussion on the approximation to the calculation of the degrees of freedom.

Shape of the Natural Residual Distributions

It is evident that the smallest possible value for any z_{ni}^2 is 0. For example, consider a dichotomously scored item. As the person's proficiency increases relative to an item's difficulty, so the expected value becomes closer and closer to 1. If the person's response is 1, then the residual will be close to zero. However, if the response is 0, then the residual will be large. The residual value has a lower bound. This will occur when the observed and expected values are the same. However, it has no upper bound—as the observed and expected values become more different, then the standardized residual increases in value, and therefore, so does z_{ni}^2. Figure 15.1 shows these possible values for the squared standardized residuals for person locations between -3 and $+3$ logits and for an item with difficulty 0.5, for both a response of 1 and a response of 0. It is clear that they can take on values with a pattern, which is formally called a *locus*. You may wish to choose a person location for an item and calculate z_{ni}^2 for an item of difficulty 0.5 and verify Fig. 15.1.

Because z_{ni}^2 has a minimum value of 0, the minimum values of y_n^2 and y_i^2 are also 0. As a result, Z_n and Z_i tend to be skewed. This skew can be ameliorated in general by an alternative transformation, which we now describe. However, it is stressed that these are all approximations and to take note of the cautions below on interpreting fit statistics. Fit statistics should be interpreted relatively, in context and from the perspective of outliers, and not against an absolute value.

Instead of forming the differences, $y_n^2 - E[y_n^2]$ and $y_i^2 - E[y_i^2]$, we form the ratios

$$y_n^2 / E[y_n^2] \tag{15.10}$$

and

$$y_i^2 / E[y_i^2]. \tag{15.11}$$

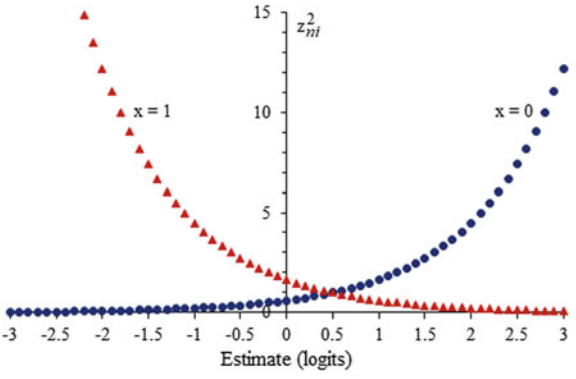

Fig. 15.1 Squared standardized residuals for a dichotomous item of difficulty 0.5

Further description of the transformation based on the ratios of Eqs. (15.10) and (15.11) is provided in the RUMM2030 Interpreting Manual (Andrich et al., 2018). The final statistic is called the *fit-residual*. This is the residual reported in the various statistics. Smith (2002) has carried out many studies on statistics related to the fit-residual index of fit.

However, RUMM2030 also shows the distribution of the residuals from Eqs. (15.8) and (15.9), and these are called *natural residual* distributions. The graphical displays can be helpful in understanding the interpretation of these residuals. These can be obtained from the main display of RUMM2030 under the heading *Residual Statistics Distribution*.

Interpreting the Sign of the Fit-Residual

Each item and each person has a summary statistic that is termed a *fit-residual* calculated as described above. In the case of a dichotomous item, the smallest set of residuals occurs when the persons have a Guttman pattern. In that case, $\sum_{n=1}^{N} z_{ni}^2$ is a minimum, and following the transformation, the value of Eq. (15.9) is negative. On the other hand, the maximum value of $\sum_{n=1}^{N} z_{ni}^2$ will occur when the response pattern is exactly the opposite of a Guttman pattern. In that case, the value of Eq. (15.9) is positive. Thus, a value that is negative and large in magnitude reflects an item with a response pattern whose empirical discrimination tends to be greater than that of the summary discrimination of the rest of the items (the ICC). Likewise, a positive value that is large in magnitude reflects an item with a response pattern whose empirical discrimination tends to be less than that of the summary discrimination of the rest of the items. The same interpretation of the sign can be made with respect to the response pattern of a person.

Outfit as a Statistic

In various software, there is a fit statistic referred to as *Outfit*. It is analogous and will give similar results and interpretations, as the fit-residual in RUMM2030.

Infit as a Statistic

There is a complementary statistic called *Infit*. This statistic is constructed in a similar way, but it weights the standardized residual for a person to an item by its variance. This means that a standardized residual where a person's location is very different from the location of an item is weighted less than one that is close to the person's location. The rationale for this weighting is that when a person is far from the item's

location, then the residual is very large for an unexpected response, as shown in Fig. 15.1, but that the effect on the estimate is less than that of an item that is close to the person's location.

The Infit statistic will generally, except in very unusual cases, show less misfit than the Outfit statistic.

The Correlation Among Residuals

If the data fit the model, then over a reasonably large number of persons and items, say 400 persons and 20 dichotomous items, the residuals should not be correlated with each other. Therefore, their correlations should be close to 0. RUMM2030 has a facility to show these correlations. In fact, if the data fit the model, then the smaller the number of items, the more these residuals will in theory show a negative correlation. There will likely be a small negative correlation among the residuals, for example of the order of −0.03. However, if the correlation is relatively large, for example, + 0.3, then this suggests that the pair of items assess some aspect, which is different from the common variable among the items.

The Principal Component Analysis (PCA) of Residuals

The Principal Component Analysis (PCA) in RUMM2030 is analogous to a factor analysis, and the results are interpreted in the same way. The hypothesis being tested with a Rasch model analysis is that the response structure is unidimensional and that, apart from a single variable and the item parameters mapped on this variable, the remaining variation is random. The PCA focuses the pattern of residuals on to successive components to summarize which subsets of items assess an aspect in common, which is not accounted for by the single variable. For example, the first principal component of the residuals might show a number of items with large positive correlations (often called loadings), and another number with negative loadings. In that case, it might be that those two sets of items assess an aspect in common which is different from what all the items assess in common.

Generally, only the first two principal components can be interpreted meaningfully in terms of subsets of items that might have aspects more in common with each other than with the rest of the items. This does not mean that such a subset of items does not assess the same variable as the other items; it simply means that these items assess an aspect, in addition, that is common among them.

The PCA can show a pattern when the individual correlations would not suggest that there are any patterns—it concentrates the evidence of any relationships.

General Principles in Assessing Fit

There are many ways of examining whether the data fit the model. No single fit statistic alone is sufficient to make a judgment on whether an item fits the model. Smith and Plackner (2009) suggest using a 'family' of fit statistics in assessing fit.

Interpreting Fit Statistics Relatively and in Context

Although the various fit statistics are constructed with a sound rationale, there are a number of reasons why they have to be used in context. First, in any analysis, there may be many items and many persons, and tests of fit made with respect to different hypotheses for misfit. And all of them are related to each other. For example, if one item is removed from a set, the fit statistics will change for all the items. It may be that only one item misfits the model, and that all others do fit. In that case, removing an item will result in the other items showing fit. However, with real data that is rare—there are degrees of fit and misfit.

In the extreme, there are items that really do misfit, and there are data sets as a whole where there is a poor fit. However, in general, fit needs to be seen as relative. Thus, fit statistics of any kind should be ordered in the first instance and those items with the worst fit studied first. Further, relatively large misfit needs to be considered an anomaly, and studied and understood substantively.

Second, the statistics involve discrete responses, and the statistics that are constructed as random normal deviates are approximations. This feature interacts with the one above to make the distributions not fit the normal distribution, when the data fit the model, perfectly.

Third, various statistics are affected by the sample size, but these are affected differently by the sample size. They are also affected by the alignment between the persons and the items—the better they are aligned, the more likely to detect misfit.

No fit statistic is necessary and sufficient to assess that the data fit the model—the misfit of an item, which looks large, may even be localized in one area of the continuum. For these reasons, the study of fit is a forensic analysis, and every data set should be considered as if it is a case study. All the relevant evidence needs to be considered in making a decision. Of course, there will be cases with such large misfit that it is clear that the item does not belong to the set. However, it is still helpful to consider this an anomaly and to understand why there is such a large misfit of an item, which presumably is in the set because the constructors considered that it should belong to the set.

Further, every conclusion should be considered against the substantive variable, its purpose and its application. It is not a matter of assessing statistical fit only against some theoretical values. Theoretical values, such as a fit-residual greater than $+3$ or -3 can be helpful, but it should not be the sole basis used for excluding an item from a test, for example. Graphs of ICC curves should also be considered.

All these fit statistics should be used as guides, and multiple pieces of evidence should be used in making any decision to modify, discard, or deal with an item in any way. Remember, the item was there because it was believed that it conformed to the theoretical construct that was being operationalized using an instrument.

Power of the Tests of Fit as a Function of the Sample Size

No real data set fits any model perfectly. Therefore, a level of precision can be found in which any model will be rejected. The precision, or in statistical terms, the power of the test of fit, is governed by the sample size. The greater the sample size, the greater the power in detecting misfit. For very large sample sizes, the power to detect misfit is so great that any data set will misfit. Therefore, some realistic sample sizes need to be used in studying fit. One guideline is that between 10 and 20 persons for every threshold in the item set should be adequate to conduct the tests of fit. Of course, if the data fit with respect to a bigger number, then that is fine.

Generally, the number of persons one has in a sample is not a function of experimental design, but how many persons need to be assessed. There is no reason why the power of the fit analysis should be directly related to the number of persons that have to be assessed.

Sample Size in Relation to the Number of Item Thresholds

Tests of fit are affected by many factors in context, including the sample size. In general, the greater the sample size, the more powerful the test of fit that the responses do not fit the model. Essentially, it is one of precision—the greater the sample size, the greater the precision of the estimates and therefore the greater the evidence if the responses do not fit the model. This is as it should be, though some fit statistics are more sensitive than others.

Often in real data, the sample size is just the size that is in the population to be tested, and it might be over a hundred thousand. In that case, no real data will fit the model. However, the precision implied by such a sample size is generally much greater than is required, and the responses give meaningful comparisons. That is, for the item calibration and fit stage, it is not necessary to have such a large sample.

One rule of thumb based on substantial experience and simulation is that the number of persons chosen should increase as the number of item thresholds increases. The ratio that seems reasonable is between 10 and 20 persons for each threshold. For example, in a dichotomous test of 20 items, it would be useful to have at least 400 persons, and in an assessment with 10 polytomous items with 3 thresholds each (30 thresholds altogether), it would be useful to have at least 600 persons. However, this does not mean that smaller numbers of people might not give meaningful results.

There may be perspectives from which the results are very interpretable, for example a very anomalous item.

Furthermore, we would not expect very meaningful values of the above fit statistics unless there were something like 20 thresholds so that there were 21 score points, that is, scores between 0 and 20, and these had reasonable frequencies. Less score points than this make the spread of the persons rather narrow. Complementing these recommendations is that the responses should not have floor or ceiling effects, that is, that the persons are not too far to one or the other end of the scale such that the items are not distinguishing among persons. Some data sets cannot avoid such effects, for example, when a clinical population is assessed with an instrument constructed to distinguish among members of a non-clinical population, or when a standard population is assessed with an instrument constructed to distinguish among members of a clinical population. However, then even greater caution is required to interpret the fit statistics. Ideally, such data should not be used to investigate the fit properties of the items.

Adjusting the Sample Size

RUMM2030 has an option for modifying the sample size for the Chi-square (χ^2) fit statistic. Table 15.1 shows the summary fit statistics and the second set of χ^2 fit statistics. The latter set involves the same data as the first one, but the sample size in the calculation has been adjusted to 1000. As a result of this adjustment, we expect the fit to appear better. As is evident, the fit is better—every item has a smaller χ^2 value and a greater probability of arising by chance.

The above adjustment is algebraic; it assumes that the observed and expected values are the same in each class interval, but that the sample size is smaller. It is possible to adjust the sample size to be larger.

This is not the same as taking a smaller random sample and rerunning the whole analysis on this random sample. Although the fit will generally be better with a smaller sample, even in this case, there is more random variation than simply adjusting the sample size in the statistic.

Power of Tests of Fit as a Function of the Separation Index

Although our estimates, when the data fit the model, are independent of the distribution of the persons, the tests of fit are not. It is necessary to have a range of person locations in order for there to be some power in the test of fit. This is indicated qualitatively with the person separation index. The greater the separation index, the greater the spread of persons relative to the standard errors, and therefore the greater the power of the test of fit. Remember, this is the power to detect misfit. The greater the

Table 15.1 Summary item fit table for 2000 persons (1900) with no extremes, and χ^2 adjusted to a sample of 1000

Sample size of 2000 (1900)								Adjusted sample size of 1000					
Item	Location	SE	FitResid	DF	Chi-Sq	DF	Prob	SE	FitResid	DF	ChiSq	DF	Prob
I0001	−0.539	0.033	−1.712	1659.63	22.446	9	0.0076	0.033	−1.712	1659.63	11.814	9	0.2240
I0002	−0.401	0.039	0.408	1659.63	10.035	9	0.3476	0.039	0.408	1659.63	5.282	9	0.8091
I0003	0.074	0.033	−0.621	1659.63	9.716	9	0.3740	0.033	−0.621	1659.63	5.114	9	0.8243
I0004	−0.122	0.037	1.164	1659.63	4.408	9	0.8825	0.037	1.164	1659.63	2.320	9	0.9853
I0005	0.038	0.037	0.459	1659.63	11.601	9	0.2367	0.037	0.459	1659.63	6.106	9	0.7293
I0006	0.238	0.036	−0.488	1659.63	13.688	9	0.1339	0.036	−0.488	1659.63	7.204	9	0.6159
I0007	0.285	0.035	−1.514	1659.63	22.181	9	0.0083	0.035	−1.514	1659.63	11.674	9	0.2323
I0008	0.426	0.032	−0.793	1659.63	17.389	9	0.0430	0.032	−0.793	1659.63	9.152	9	0.4233

distance of the majority of persons from an item, the greater the power of detecting misfit for that item.

Test of Fit is Relative to the Group and the Set of Items

In particular, every test of fit of an item is relative to the total set of items. Thus, the Rasch model estimates the parameters in the model from all of the items. In principle, the model parameters are those that come from the data on the assumption of the model. If an item is removed, then the rest of the items provide the frame of reference for the estimates and fit. The fit values will change if an item is deleted. If one item shows a large misfit compared to the others, then if this item is removed the rest might fit well.

Bonferroni Correction

Typically, many tests of fit are conducted. There is concern that with many tests of fit, some will be significant just by chance. There are suggestions for correction of the significance level in the literature, and a common one is the Bonferroni correction (Bland & Altman, 1995). This is very simple to carry out—the chosen probability value of significance is simply divided by the number of tests of fit. The Bonferroni correction is an adjustment to the significance level to reduce the risk of a type I error. A type I error occurs when a significant misfit is found when there is none. A type II error occurs when no misfit is found when there is one. There is some controversy with this correction. In RUMM2030, both the numbers with correction and the numbers without correction are provided to give the users discretion in making decisions. It also permits them to report both.

RUMM2030 Specifics

In RUMM2030, the χ^2 and item fit-residual statistics are provided in addition to graphical evidence of item fit, the item ICCs. There is also an option to calculate and display another fit statistic, based on ANOVA (*Include ANOVA Item Fit Statistics* checkbox on the Analysis Control form). There are no absolute criteria for interpreting fit statistics. The default for the fit-residual statistic in RUMM2030 is 2.5 but that can be changed (Change *Residual criterion* on the Analysis Control form).

A total item–trait interaction χ^2 statistic is provided with its probability value and degrees of freedom, which is the number of items multiplied by the item degrees of freedom (χ^2 degrees of freedom for an item is the number of class intervals minus 1). The total item–trait interaction χ^2 statistic reflects the property of invariance across

the trait. A significant value means that the hierarchical ordering of one or more items varies across the trait. Also provided are the item fit-residual mean and SD, with ideal values of 0 and 1, respectively.

The default number of class intervals will be 10 for a sample of N = 1000. The initial number of class intervals is calculated by RUMM2030 to have at least 50 persons in a class interval, if possible. There is an option to change the *Number of Class Intervals* on the Analysis Control form (the minimum number is 2 and the maximum is 10). RUMM2030 allocates persons to class intervals based on the person location distribution for the total sample. If missing data is present, then some or not all persons will have responded to every item, possibly leading to very small numbers of persons in specific class intervals for some items. In this case, the class interval distributions are adjusted on an item-by-item basis in RUMM2030.

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.
Exercise 6: Analysis of data with dependence in Appendix C.

References

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.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *British Medical Journal, 310,* 170.

Smith, E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3,* 205–231.

Smith, R. M., & Plackner, C. (2009). The family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement, 10*(4), 423–437.