Chapter 5

Item Fit Statistics

5.1. Introduction

Items from Rasch models and items from other IRT models share a number of properties, and differ only with respect to the choice of functions describing how the probabilities of responses to items depend on the latent variable. The properties shared by all IRT models are unidimensionality, monotonicity, local independence and no differential item functioning. These properties are often motivated by subject matter theory about the latent variable that items are supposed to measure. The same can rarely be said about the choice of the specific probability function of the Rasch model and it is often argued that the parsimonious Rasch model is too simple to have a chance to fit real-life data. Therefore, it is important to provide strong empirical evidence supporting the claim that the Rasch model is adequate for data.

This chapter describes two types of item fit statistics that can be used for this purpose. The first type takes all the fundamental assumptions of the Rasch model as given and tries to assess the degree to which the separate items appear to have conditional response probabilities that do not depart from the Rasch model probabilities. The second type addresses the assumption of no differential item functioning, but does it one item at a time, assuming all the other items do not violate the Rasch model assumptions.

The methods described in this chapter always target separate items. It is therefore common that evidence against the data-model fit provided by these methods is often taken for evidence against specific items that do not fit the model, rather than against the fit of the model to data. The degree to which such a viewpoint

---

Chapter written by Karl Bang CHRISTENSEN and Svend KREINER.

is viable is controversial. We refer to Andrich [AND 04] for a discussion of these issues. What we want to stress here is that violations of the fundamental assumptions of unidimensionality, local independence and no DIF imply that the idea that the separate items should follow a Rasch model (or any other standard IRT model sharing these assumptions) looses its meaning. Before we dismiss an item because of its disagreement with the response probabilities of the model, we should always check the assumptions of unidimensionality, local response dependence and DIF among the other items. We will return to these issues at the end of the chapter.

Item response theory, in general, and Rasch models, in particular, contain a plethora of different item fit statistics. It is beyond the scope of this book to discuss these in detail, but we will discuss a number of item fit statistics that are useful and intuitive. The interested reader should refer to Glas and Verhelst [GLA 95a, GLA 95b] or Smith [SMI 04] for an overview.

There are a number of misunderstandings concerning how statistical methods work. For this reason, a number of concrete ideas for testing the fit of an item to the Rasch model have been implemented in ways that may produce erroneous results during the analysis. To make it easier to assess these issues, we discuss the Rasch model residuals, which compare observed responses with expected responses, and the way they should be used to test the fit of items to Rasch models.

5.2. Rasch model residuals

In statistics, residuals are defined as the difference between observed data and expected values under a specific model and/or hypothesis. The theory of Rasch models describes and uses two types of residuals: individual response residuals and group residuals. To clearly distinguish between the two types of residuals, we need first to introduce some notations.

5.2.1. *Notation*

Let X_{vi} denote the response of person v to item i and let $E_{vi} = E(X_{vi})$ denote its expectation. Under the Rasch model, E_{vi} is given by

$$E_{vi} = \sum_{x=0}^{m_i} x \frac{\exp(x\theta_v + \psi_{ix})}{K(\theta_v, \psi_{i0}, \ldots, \psi_{im_i})} \qquad [5.1]$$

Because item and person parameters are unknown, it is common to use estimates of the parameters instead of the true parameter. This leads to *estimates* of expected item scores given by

$$\hat{E}_{vi} = \sum_{x=0}^{m_i} x \frac{\exp(x\hat{\theta}_v + \hat{\psi}_{ix})}{K(\hat{\theta}_v, \hat{\psi}_{i0}, \ldots, \hat{\psi}_{im_i})} \quad [5.2]$$

We use these when we want to make it clear that expected item scores are based on parameter estimates rather than known parameters.

Let $R_v = \sum_{i=1}^{k} X_{vi}$ denote the total score over all items. The maximum score over all items is $m_\cdot = \sum_{i=1}^{k} m_i$. Let G_v denote a categorical variable with m_G categories. Finally, let S_v be an ordinal categorical variable defined by m_S score intervals.

Let n be the number of persons in the complete sample. To define group residuals, we refer to the number of persons in different groups. We do this in the following way: n_r denotes the number of persons with $R_v = r$, n_g denotes the number of persons with $G_v = g$, n_s denotes the number of persons with $S_v = s$, and n_{gr} denotes the number of persons with $G_v = g$ and $R_v = r$. Unless specifically stated, we always assume that persons with extreme scores (0 and $m_\cdot$) are excluded from these counts because there is no information in the response profiles of these persons that are useful during the analysis of the fit of the items to the Rasch model.

At the group level, we will deal with residuals that compare observed and expected average item scores in groups. We therefore define average item scores in groups defined by exogenous variables

$$A_{gi} = \frac{1}{n_g} \sum_{v:G_v=g} X_{vi} \quad [5.3]$$

by score intervals

$$A_{si} = \frac{1}{n_s} \sum_{v:S_v=s} X_{vi} \quad [5.4]$$

and by raw scores

$$A_{ri} = \frac{1}{n_t} \sum_{v:R_v=r} X_{vi} \quad [5.5]$$

Group residuals compare the observed average scores [5.3]–[5.5] to the expected values $B_{gi} = \frac{1}{n_g} \sum_{v:G_v=g} E_{vi}$, $B_{si} = \frac{1}{n_s} \sum_{v:S_v=s} E_{vi}$ and $B_{ri} = \frac{1}{n_r} \sum_{v:R_v=r} E_{vi}$.

5.2.2. *Individual response residuals: outfits and infits*

The raw residuals are

$$R_{vi} = X_{vi} - E_{vi}$$

Because item and person parameters are unknown, it is a standard statistical practice to use $\hat{E}_{vi}$ instead of E_{vi} for the calculation of residuals. The problems created by this practice during item analysis by Rasch models are discussed below, but for now we disregard these problems and assume that the true expected values are available and define individual response residual statistics in the following way. The standardized residuals are

$$Z_{vi} = \frac{X_{vi} - E_{vi}}{\sqrt{VAR(X_{vi} - E_{vi})}} = \frac{R_{vi}}{\sqrt{VAR(X_{vi})}} \quad [5.6]$$

The squared standardized residuals are

$$Z_{vi}^2 = \frac{R_{vi}^2}{VAR(X_{vi})} \quad [5.7]$$

The theory of Rasch models defines two summary item fit statistics called OUTFIT and INFIT based on these residuals:

$$\text{OUTFIT}_i = \frac{1}{n}\sum_{v=1}^{n} Z_{vi}^2 \quad [5.8]$$

$$\text{INFIT}_i = \frac{\sum_{v=1}^{n} R_{vi}^2}{\sum_{v=1}^{n} VAR(X_{vi})} \quad [5.9]$$

Because $R_{vi}^2 = VAR(X_{vi})Z_{vi}^2$, we can redefine the infit as a weighted mean of the standardized squared residuals Z_{vi}^2:

$$\text{INFIT}_i = \sum_{v=1}^{n} w_{wi} Z_{vi}^2 \quad [5.10]$$

where $w_{vi} = VAR(X_{vi}) / \sum_{v=1}^{n} VAR(X_{vi})$.

The range of these statistics consists of non-negative real numbers and they both have expected values equal to one under the Rasch model. Infit or outfit values close to zero or much greater than one are therefore regarded as evidence against the fit of an item to the Rasch model. Exactly when these statistics are too small or too large to be acceptable is, however, a difficult question and the established practice surrounding these fit statistics is infested with a number of misunderstandings and misconceptions. In this chapter, we therefore take a closer look at how to calculate the expected responses when item and person parameters are estimated values and on the distributions of the item outfit and infit under the Rasch model.

5.2.3. *Problem 1: the distribution of outfit and infit test statistics*

Let us take the distribution first. We need to know the distribution of outfit and infit test statistics if we want to assess whether or not the observed outfit and infit test statistics disagree with the Rasch model. It is sometimes claimed [SMI 04] that the standardized residuals Z_{vi} "are distributed as approximate unit normals" and thus Z_{vi}^2 "can be evaluated as a chi-square (χ^2) with one degree of freedom", so that the "sum across the persons to assess the fit of the responses to a particular item ... results in a chi-square".

Unfortunately, these claims are false. This is not a new result [VAN 82], but belief in these claims persists, and widely used computer programs for item analysis by Rasch models implement assessment of the significance of the fit statistics based on these misunderstandings.

It is not difficult to see why the claims are wrong. Recall that item responses X_{vi} are categorical variables with a relatively small number of possible outcomes. Subtracting the expected response from X_{vi} does not change this and the resulting residual statistics, R_{vi}, Z_{vi} and Z_{vi}^2, are therefore also discrete with the same number of potential outcomes as X_{vi}. The standardized residual of a binary variable is, for instance, still a binary variable and therefore as far away from the continuous normal distribution as we can imagine.

We cannot assume that the Z_{vi} variables are identically distributed but the means and variances of these variables are the same. We might therefore hope for a version of the central limit theorem (CLT) to guarantee that the asymptotic distribution of the average of the Z_{vi} variables was normal. Such CLT versions do exist, but they require some additional assumptions concerning the distribution of the Z_{vi} variables are satisfied and it is not clear at all whether the Z_{vi} variables meet these requirements. Transforming the standardized residuals into squared standardized residuals complicates things to a degree where even this hope is unrealistic. It is true that the expected values of Z_{vi}^2 are equal to one, but its standard deviation (which would equal to two if Z_{vi}^2 had a χ^2 distribution with one degree of freedom) depends strongly on E_{vi}. Consider, as an example, two dichotomous items with E_{vi} equal to 0.5 and 0.1, respectively. In the first case, Z_{vi} can be either -1 or $+1$ with probabilities

$$P(Z_{vi} = -1) = P(Z_{vi} = +1) = 0.5$$

so that Z_{vi}^2 always is equal to one, $P(Z_{vi}^2 = 1) = 1$, and $VAR(Z_{vi}^2) = 0$. In the second case, Z_{vi}^2 has one of the two values, $P(Z_{vi}^2 = 0.1111) = 0.9$ and $P(Z_{vi}^2 = 9) = 0.1$ and $VAR(Z_{vi}^2) = 7.1$. Apart from illustrating that $VAR(Z_{vi}^2)$ is very different from two, the example shows that the result of the test of fit of an item would depend strongly on the targeting of the items to the population. If the set of items is well targeted with many persons with probabilities of positive responses

that were relatively close to 0.5, it would be extraordinary to find items that did not fit the Rasch model. If items are off target with many persons with low probabilities of positive responses on many items, the opposite would be the case.

According to formula [5.1], there is one expected item score and therefore one residual distribution for each combination of a person and an item. If this statement is taken at face value, there is no easy solution to the problem of finding an appropriate distribution for the outfit and infit test statistics. There is, however, every reason to take a closer look at the way to calculate E_{vi} when person and item parameter estimates are used instead of the unknown parameters, and this will not only lead to better ways to calculate E_{vi} but also to a more viable solution to the distribution problem.

5.2.4. *Problem 2: calculating* E_{vi}

It is easy to find examples – linear regression analysis being the most obvious – of statistical analyses where $\hat{E}_{vi}$ can replace E_{vi} during the analysis of residuals. The reason why this often works in practice is that the analyses use consistent estimates of unknown parameters where bias and standard errors disappear as sample sizes increase toward infinity. These situations are, however, very different from the situation with the individual response residuals in Rasch models. The problem is that $\hat{E}_{vi}$ depend on two types of parameters: item parameters and person parameters where consistent estimates are available for one, but not for both types of parameters. Item parameter estimates can be assumed to be consistent except for the so-called joint estimates that are known to be inconsistent. Person parameter estimates could also be regarded as consistent if the number of items is large, but this is never (or at least very rarely) the case in health-related scales where the typical number of items lies between five and 25.

Recall the discussion of person parameter estimates in Chapter 4 where it has been shown that standard errors can be large for all persons and the bias can be considerable for a large number of persons. If there are no other reasons, it follows that *estimates* of expected values must be expected to be biased with relatively large standard errors. Recall also that the bias of the person parameter estimate depends on the choice of an estimator. Before we calculate $\hat{E}_{vi}$, we have to think carefully about the choice of person parameter estimate: ML, WML, JML or Bayesian. The estimates can be very different – some with more bias and error than others – but since we want to estimate $\hat{E}_{vi}$, we have to select the person parameter estimate that reduces the error and the bias of $\hat{E}_{vi}$, and there are – to our knowledge – no published results about the degree to which the differences between person parameter estimates influence the error and the bias of $\hat{E}_{vi}$. These problems are in themselves reason enough for a cautious attitude toward the application of $\hat{E}_{vi}$, but there is another problem that we also have to take into account: formula [5.2] does not give the correct value because estimates of person locations are used instead of the true values. To see this, we have to take a closer look at what happens during the calculation of the expected values.

Recall first that the Rasch model probabilities are *conditional* probabilities given outcomes on a latent variable. The proper way to write these probabilities is

$$P(X_{vi} = x|\Theta_v = \theta; \bar{\psi}_i) = \frac{\exp(x\theta + \psi_{ix})}{K(\theta; \bar{\psi}_i)} \quad [5.11]$$

where Θ_v is the latent variable, θ is an outcome on this variable and $\bar{\psi}_i$ is the vector of item parameters. Attempting to replace θ by an estimate $\hat{\theta}$ in [5.11] means that we replace the variable Θ_v in [5.11] by a different variable $\hat{\Theta}_v$ with outcomes equal to the estimates $\hat{\theta}$, thus replacing [5.11] by

$$P(X_{vi} = x|\hat{\Theta}_v = \hat{\theta}; \bar{\psi}_i) = \frac{\exp(x\hat{\theta} + \psi_{ix})}{K(\hat{\theta}; \bar{\psi}_i)} \quad [5.12]$$

Next, recall that the total score R is sufficient for θ. From this, it follows that the range of $\hat{\Theta}_v$ consists of $m_\cdot + 1$ values

$$\{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(r)}, \ldots, \theta^{(m_\cdot)}\}$$

where $\hat{\Theta}_v = \theta^{(r)}$ if and only if $R_v = r$. Furthermore, the true conditional probability $P(X_{vi} = x|\hat{\Theta}_v = \hat{\theta}; \bar{\psi}_i)$ cannot be equal to the probability in formula [5.12]. The true probability has to be derived from the conditional distribution of X_{vi} given R_v:

$$\begin{aligned} P(X_{vi} = x|\hat{\Theta}_v = \theta_r; \bar{\psi}_i) &= P(X_{vi} = x|R_v = r; \bar{\psi}_i) \\ &= \frac{\exp(\psi_{ix})\gamma_{r-x}\left(\bar{\psi}_1, \ldots, \bar{\psi}_{i-1}, \bar{\psi}_{i+1}, \ldots, \bar{\psi}_k\right)}{\gamma_r(\bar{\psi})} \end{aligned} \quad [5.13]$$

Formula [5.13] defines the correct conditional distribution of X_{vi} given $\hat{\Theta}_v = \theta^{(r)}$ and illustrates the advantages of conditional inference in Rasch models. According to formula [5.13], the conditional probability of X_{vi} given $\hat{\Theta}_v = \theta^{(r)}$ does not depend on the *value* of $\theta^{(r)}$, but only on r. This means that (1) we do not have to be concerned about how person parameters are estimated, (2) $E_{vi} = E(X_{vi}|R_v = r)$ when $\hat{\Theta}_v = \theta^{(r)}$, and (3) formula [5.13] also can be used to derive the conditional probabilities of R_{vi}, Z_{vi}, R_{vi}^2 and Z_{vi}^2 because these variables are simple functions of X_{vi} and E_{vi}. Finally, formula [5.13] shows that the distributions of X_{vi}, R_{vi}, Z_{vi}, R_{vi}^2 and Z_{vi}^2 are the same for all persons with the same score (and therefore the same person parameter estimate) so that the CLT can be used to determine the asymptotic distribution of outfit and infit test statistics.

For instance, the outfit test statistic defined in [5.8] can be rewritten as a weighted mean of the average of a number of identically distributed Z_{vi}^2 values

$$\text{OUTFIT}_i = \frac{1}{n}\sum_{v=1}^{n} Z_{vi}^2 = \frac{1}{n}\sum_{r=1}^{m_\cdot - 1}\sum_{v:R_v=r} Z_{vi}^2 = \sum_{r=1}^{m_\cdot - 1} w_r V_{ri} \quad [5.14]$$

where $w_r = \frac{n_r}{n}$ and

$$V_{ri} = \frac{1}{n_r} \sum_{v:R_v=r} Z_{vi}^2$$

is the average of n_r identically distributed outcomes on Z_{vi}^2. It follows from the CLT that V_r has an asymptotically normal distribution whose mean is equal to one and whose variance depends on the conditional variance of Z_{vi}^2 given $R_v = r$. Since OUTFIT_i is a weighted mean of $m_{\cdot} - 1$ variables with asymptotically normal distributions, it follows that OUTFIT_i is also asymptotically normal under the Rasch model with a mean that is equal to one and with a variance that is a weighted mean of the variances:

$$VAR(\text{OUTFIT}_i) = \sum_{r=1}^{m_{\cdot}-1} w_r^2 VAR(V_{ri}) \qquad [5.15]$$

The asymptotic distribution of the item INFIT defined in [5.9] and [5.10] can be derived in the same way.

EXAMPLE 5.1.– We continue the analysis of the items from the disinhibited eating (DE) subscale of the Diabetes Health Profile (DHP) with calculation of outfit and infit tests statistics. Table 5.1 shows the results.

	OUTFIT			INFIT		
Item	Observed	SD	p-value	Observed	SD	p-value
DHP32	1.237	0.107	0.027	1.246	0.108	0.022
DHP34	1.042	0.146	0.772	0.910	0.107	0.400
DHP36	1.260	0.110	0.018	1.190	0.094	0.043
DHP38	0.754	0.125	0.048	0.815	0.128	0.147
DHP39	0.937	0.128	0.624	0.923	0.119	0.517

Table 5.1. *Outfit and infit test statistics for the five items in disinhibited eating (DE) subscale of the Diabetes Health Profile (DHP)*

Significant evidence of misfit was found for three out of five items. However, the evidence provided by p-values greater than 1% is not strong, and adjustment for multiple testing by, for example, the Benjamini–Hochberg procedure [BEN 95] dismisses this evidence. So according to these fit statistics, the DHP–DE items do fit the Rasch model.

5.2.5. *Group residuals*

Group residuals compare the *average* item score in a group A_{gi} to the expected *average* score B_{vi} under the Rasch model. There is a close link between individual

residuals and group residuals, but there are also important differences. Before we go into details, it should be noted that the introduction of groups to some extent changes the issue that the item fit statistics try to address. Groups can be defined in three different ways during the analysis of the data-model fit: (1) score groups, (2) score intervals and (3) groups defined by exogenous variables.

The natural groups under Rasch models are the score groups consisting of persons with the same total score because all persons in these groups have the same conditional response probabilities for all items. For large sample sizes, item fit within these groups should be analyzed separately. When this is not the case, we can define groups by score *intervals* instead of separate scores. This complicates things in technical terms, but in principle, there is no difference between the analysis using score groups and the analysis using score intervals. Since low and high scores are associated with low and high values of θ, we could say that the analysis of group residuals defined by the scores on all items tries to do the same as item fit statistics based on individual response residuals, namely testing that items are homogeneous in the sense that they function in the same way at all levels of θ. Groups may also be defined by exogenous variables that are not part of the Rasch model as such. When this happens, analyses of residuals address a completely different and, to some extent, a more fundamental issue, namely the question of whether items function in the same way in different groups. If the groups, for instance, are defined by gender, we summarize residuals calculated for men and women separately to test that there is no differential item functioning (DIF) relative to gender.

Many of the methods used during the analysis of homogeneity are the same as those used during the analysis of DIF, but there are also important differences where methods do not transfer effortlessly from one type of problem to the other. Therefore, we consider the analysis of homogeneity first, and return to a discussion of the analysis of DIF later.

5.2.6. *Group residuals for analysis of homogeneity*

It is easy to see that the group residual is nothing but the average individual residual in the group

$$D_{gi} = A_{gi} - B_{gi} = \frac{1}{n_g} \sum_{v:G_v=g} X_{vi} - \frac{1}{n_g} \sum_{v:G_v=g} E_{vi} = \frac{1}{n_g} \sum_{v:G_v=g} R_{vi} \quad [5.16]$$

The standardized group residual is

$$F_{gi} = \frac{D_{gi}}{\sqrt{VAR(A_{gi})}}$$

where

$$VAR(A_{gi}) = \frac{1}{n_g^2} \sum_{v:G_v=g} VAR(X_{vi})$$

and the squared standardized group residual is

$$F_{gi}^2 = \frac{\left(\sum_{v:G_v=g} R_{vi}\right)^2}{\sum_{v:G_v=g} VAR(X_{vi})} \qquad [5.17]$$

We still assume that E_{vi} is the *conditional* mean of X_{vi}, given $R_v = r$ so that A_{gi} is a weighted mean of the averages of item responses in a number of score groups and that A_{gi} is asymptotically normal. Thus, the asymptotic distribution of F_{gi} is standardized normal and the asymptotic distribution of F_{gi}^2 is a χ^2 distribution with one degree of freedom.

Therefore, we might expect that inference summarizing information on a specific item from several groups should be easier, but this is not the case as illustrated in the following example.

EXAMPLE 5.2.– Consider the results in Table 5.2 concerning group residuals in the score intervals from one to five and six to 15. Each of these score intervals provides evidence for or against the fit of the items to the Rasch model. The item fit statistics appear to agree in the majority of cases; in most cases, the evidence from the different score intervals appears to agree, but there is one item (DHP38) where the conclusions differ and there is another item (DHP32) where the evidence from the two score intervals is insignificant taken separately but where we might expect that an assessment based on a summary of the two residuals also might result in evidence against the item-model fit.

Score 1–5				
		Mean		
Item	n	Obs.	Exp.	F_{gi}
DHP32	85	0.61	0.50	1.73
DHP34	85	0.66	0.71	−0.61
DHP36	85	0.81	0.72	1.07
DHP38	85	0.29	0.43	−2.35
DHP39	85	0.88	0.89	−0.15

Score 6–14				
		Mean		
Item	n	Obs.	Exp.	F_{gi}
DHP32	91	1.13	1.23	−1.29
DHP34	91	2.41	2.36	0.59
DHP36	91	1.79	1.88	−1.00
DHP38	91	1.21	1.08	1.63
DHP39	91	1.61	1.61	0.10

Table 5.2. *Analysis of group residuals comparing observed with expected item scores in two score intervals (F_{gi} is the standardized residual)*

An obvious solution to this problem is to calculate the sum of the squared residuals and to assess the significance according to a χ^2 distribution with degrees of freedom

equal to the number of groups. Doing this would be an error similar to the error we would make if we assumed that the sum of the squared individual residuals has a χ^2 distribution, but for a different reason. The reason why one would try to calculate $\sum_{g=1}^{m_G} F_{gi}^2$ is that the CLT guarantees that F_{gi}^2 in itself has an asymptotic χ^2 distribution with one degree of freedom and that the sum of χ^2 distributed variables is known to have a χ^2 distribution. However, this result only applies when the variables are stochastically independent and this requirement is not met by the squared group residuals in Rasch models. To see that this is the case, we only have to note that the observed minus expected item scores in the low-score interval is equal to the expected minus observed item scores in the high-score group (except for rounding error). This is a general result that has to do with the sufficiency of the item margins under the Rasch model and the way item parameters are estimated, and it follows that the squared standardized residuals for the same item in different groups are correlated: we cannot assume that $\sum_{g=1}^{m_G} F_{gi}^2$ has a χ^2 distribution.

The requirement that χ^2 distributed variables have to be independent for the sum to be χ^2 distributed is often forgotten and problems caused by this can be found in many summary fit statistics that are implemented in software for Rasch analysis. We will not go deeper into this issue, but advise the reader to take a careful look at software documentation to ensure that implemented fit statistics is not flawed by this mistake.

5.3. Molenaar's U

We will not make that error here and therefore have to look at other ways to summarize the evidence against items from different groups. One of these is the U statistic proposed by Molenaar [MOL 83] summarizing trends in group residuals calculated in different score groups.

Let F_{ri} be the standardized residual for item i in score group r. Molenaar noticed that the patterns in these residuals often had negative values at one end of the score range and positive values at the other end corresponding to what we would expect if the item discrimination of an item were different from the item discrimination of the other items. To capture such departures from the Rasch model, Molenaar partitioned the score range into three intervals

$$\{1, \ldots, r_1\}, \{r_1 + 1 \ldots, r_2\}, \{r_2 + 1, \ldots, m_\cdot - 1\}$$

with approximately the same number of respondents, he defined his item fit statistic as

$$U_i = \frac{1}{\sqrt{r_1 + (m_\cdot - 1 - r_2)}} \left(\sum_{r=r_2+1}^{m_\cdot} F_{ri} - \sum_{r=1}^{r_1} F_{ri} \right) \qquad [5.18]$$

and showed that it has an asymptotic standardized normal distribution under the Rasch model. Positive or negative values indicate that the discriminations of item i is stronger or weaker, respectively, than what is assumed by the Rasch model.

EXAMPLE 5.3.– Table 5.3 shows the values of Molenaar's U for the five DHP items.

Item	U_i	p-value
DHP32	-1.892	0.059
DHP34	0.170	0.865
DHP36	-1.003	0.316
DHP38	2.026	0.043
DHP39	0.780	0.435

Table 5.3. *Molenaar's U comparing residuals in the 0–3 score interval ($n = 45$) to residuals in the 7–14 score interval ($n = 64$) for the five items from the disinhibited eating (DE) subscale of the Diabetes Health Profile (DHP)*

The value of U_i for item DHP38 is significantly greater than zero, but adjusting for multiple testing once again dismisses the significance so that the conclusion drawn from this statistic is the same as the conclusion drawn from the infit and outfit test statistics in Table 5.1: there is no conclusive evidence against the item-model fit.

Molenaar's U attempts to capture the trend in the residuals (if there is one) relative to the total score and therefore focuses on the extreme ends of the score range. The idea is sound, but we might be concerned about the power of the fit statistic because the coefficient disregards about a third of the sample of persons. In the analysis of the five items from the DE subscale of the DHP, U disregarded item responses from 57 respondents. To include these respondents in the analysis of item fit, Kreiner [KRE 11] proposed another item fit statistic targeting this kind of departures from the Rasch model. This statistic is discussed in the following section.

5.4. Analysis of item-restscore association

Instead of looking at residuals that compare *average* item scores to expected average item scores, Kreiner's fit statistic compares observed and expected response frequencies in different score groups to the expected responses in the score groups.

Recall that n_r denotes the observed number of respondents with $R_v = r$ and let $a_{rx}^{(i)}$ denote the number of respondents with $R_v = r$ and $X_{vi} = x$. Given n_r, the expected number of respondents is

$$b_{rx}^{(i)} = n_r P(X_{vi} = x | R_v = r)$$

The conditional distribution of item responses given the total score is one of the standard distributions from the theory of Rasch models (see Chapter 2). Intuition might suggest that this leads to a simple χ^2 test comparing observed to expected frequencies but this would be wrong for exactly the same reasons that summing squared residuals from different score intervals was wrong. Instead, Kreiner [KRE 11] transforms the observed and expected item by score tables into item by restscore tables.

Let $R_{vi} = R - X_{vi}$ denote the restscore without item i. The observed frequencies of the item–rest score table are

$$o_{rx}^{(i)} = a_{(r-x)x}^{(i)}$$

and the expected frequencies of the item–rest score table are

$$f_{rx}^{(i)} = e_{(r-x)x}^{(i)}$$

so that the expected table is still calculated conditionally given the total score on all items. Kreiner argues that a too strong item discrimination will result in an item–rest score correlation that is larger than expected under the Rasch model and that the result of a too weak item discrimination will be that the observed item–rest score correlation will be less than expected. To test whether this is the case, we may calculate Goodman and Kruskal's γ [GOO 54] measuring the degree of association between the item and the rest score in the observed and expected item–rest score tables together with the standard deviation of γ in the expected table and use this to test whether the observed γ disagrees with what the Rasch model expects.

EXAMPLE 5.4.– The results from an analysis of the DE items are shown in Table 5.4.

Item	Obs. γ	Exp. γ	SD	p-value
DHP32	0.305	0.426	0.068	0.07
DHP34	0.500	0.465	0.060	0.57
DHP36	0.345	0.445	0.062	0.11
DHP38	0.686	0.432	0.072	0.0004
DHP39	0.522	0.470	0.070	0.45

Table 5.4. *Analysis of the item–rest score association in five items from the disinhibited eating (DE) subscale of the Diabetes Health Profile (DHP)*

Note that item–rest score associations are not expected to be exactly the same for all items, rather they are expected to be relatively close. The observed association between DHP38 and the rest score is much stronger than expected and the difference is, according to the γ coefficient, highly significant. The item–rest score association therefore rejects the fit of DHP38 to the Rasch model.

5.5. Group residuals and analysis of DIF

The group residuals that were used for the analysis of homogeneity in Table 5.2 can also be calculated for the analysis of DIF when groups are defined by values of a covariate instead of the total score on items. This is illustrated in the following example.

EXAMPLE 5.5.– Table 5.5 shows the group residuals for men and women.

Men				
		Mean		
Item	n	Obs.	Exp.	F_{gi}
DHP32	91	0.72	0.78	-0.78
DHP34	91	1.33	1.39	-0.67
DHP36	91	1.36	1.18	2.17
DHP38	91	0.56	0.66	-1.52
DHP39	91	1.20	1.16	0.51

Women				
		Mean		
Item	n	Obs.	Exp.	F_{gi}
DHP32	84	1.04	0.98	0.78
DHP34	84	1.80	1.73	0.71
DHP36	84	1.25	1.44	-2.26
DHP38	84	0.96	0.86	1.51
DHP39	84	1.31	1.35	-0.50

Table 5.5. *Analysis of group residuals comparing observed to expected item scores among men and women. F_{gi} is the standardized residual*

In this case, the residuals in both groups agree that one item (DHP36) has a DIF problem. It still would have been convenient if it had been possible to add the squared standardized residual together to get one χ^2 statistic summarizing all the evidence concerning DIF in an item, but this is, of course, also impossible here for exactly the same reasons as before: the separate χ^2 statistics are not independent. So we have to find ways to test for DIF where responses from all groups are summarized and where the asymptotic distribution is known. Two tests are discussed: (1) a conditional likelihood ratio (CLR) test suggested by Kelderman [KEL 84] and (2) a non-parametric test of conditional independence in three-way tables. We regard the first test as the fundamental test of DIF in Rasch models, but note that the second test is very popular and widely used because it is easy to calculate. It is, however, not always recognized by users of this test that the test is a Rasch model test because the conditional independence assumptions are only true under the Rasch model.

5.6. Kelderman's conditional likelihood ratio test of no DIF

Assume that G is an exogenous variable with values $0, \ldots, m_G$ and that there is DIF in item X_i relative to G. We say that DIF is uniform within the Rasch model if the Rasch model fits in all groups defined by G and if the item parameters of all the other items are the same in all groups as illustrated below.

G	X_1	X_2	...	X_{i-1}	X_i	X_{i+1}	...	X_k
0	β_1	β_2	...	β_{i-1}	$\beta_i^{(0)}$	β_{i+1}	...	β_k
1	β_1	β_2	...	β_{i-1}	$\beta_i^{(1)}$	β_{i+1}	...	β_k
:	:	:		:	:	:		:
m_G	β_1	β_2	...	β_{i-1}	$\beta_i^{(m_G)}$	β_{i+1}	...	β_k

It is convenient to reparameterize the model by letting

$$\beta_i^{(g)} = (\beta_{i1}^{(g)}, \ldots, \beta_{im}^{(g)})$$

and

$$\beta_i = \beta_{i1}^{(0)}$$

and

$$\delta_{ix}^{(g)} = \beta_{ix}^{(g)} - \beta_{ix}^{(0)}$$

so that $\delta_{ix}^{(0)} = 0$. Replacing the previous parameters by these new parameters yields the structure shown below

G	X_1	X_2	...	X_{i-1}	X_i	X_{i+1}	...	X_k
0	β_1	β_2	...	β_{i-1}	β_i	β_{i+1}	...	β_k
1	β_1	β_2	...	β_{i-1}	$\beta_i + \delta_i^{(1)}$	β_{i+1}	...	β_k
:	:	:		:	:	:		:
m_G	β_1	β_2	...	β_{i-1}	$\beta_i + \delta_i^{(m_G)}$	β_{i+1}	...	β_k

The joint distribution of item responses is

$$P(x_1, \ldots, x_k|\theta, G = g) = \frac{\exp\left(\sum_{i=1}^{k}\left(x_i\theta - \sum_{j=1}^{x_i}\beta_{ij}\right) + \delta_{ix_i}^{(g)}\right)}{K(\theta, \beta, \delta)}$$
$$= \frac{\exp\left(r\theta - \sum_{i=1}^{k}\sum_{j=1}^{x_i}\beta_{ij} + \delta_{ix_i}^{(g)}\right)}{K(\theta, \beta, \delta)} \qquad [5.19]$$

where $\delta_{ix}^{(g)}$ are log-linear interaction parameters describing the DIF effect that G has on item i and where $r = \sum_{i=1}^{k} r_i$ is sufficient as it was in the original Rasch model. It follows from the sufficiency that conditional inference including CLR tests is also available in Rasch models with uniform DIF items. The conditional distribution of items given the total score and given G is equal to

$$P(x_1, \ldots, x_k|r, G = g) = \frac{\exp\left(-\sum_{i=1}^{k}\sum_{j=1}^{x_i}\beta_{ij} + \delta_{ix_i}^{(g)}\right)}{\gamma_r^g(\beta, \delta)} \qquad [5.20]$$

To estimate the item and DIF parameters under the uniform DIF model, we maximize the likelihood function

$$L_1(\beta, \delta) = \prod_{v=1}^{n} P(x_{v1}, \ldots, x_{vk} | r_v, g_v)$$

and calculate a CLR test comparing $L_1(\beta, \delta)$ to the likelihood of the Rasch model without DIF

$$L_0(\beta, \delta = 0) = \prod_{v=1}^{n} P(x_{v1}, \ldots, x_{vk} | r_v, g_v)$$

by

$$\chi^2 = -2 \log \left(\frac{L_0}{L_1} \right)$$

The CLR test statistic has an asymptotic χ^2 distribution with $m_G \cdot m_i$ degrees of freedom.

EXAMPLE 5.6.– The CLR tests of DIF relative to gender are shown in Table 5.6.

Item	CLR	df	p-value
DHP32	1.47	3	0.69
DHP34	3.19	3	0.36
DHP36	12.72	3	0.0053
DHP38	5.47	3	0.14
DHP39	4.93	3	0.18

Table 5.6. *CLR tests of DIF with respect to gender for the five items of the disinhibited eating (DE) subscale of the Diabetes Health Profile (DHP)*

According to these tests, DHP36 is the only item with a DIF problem and the evidence is too strong to be dismissed by adjustment for multiple testing.

5.7. Test for conditional independence in three-way tables

A completely different approach to the analysis of DIF is to test for conditional independence of an item and an exogenous variable in three-way tables where the relationship between the item and the covariate is stratified according to the score. When items are dichotomous variables and covariates are binary, such tests are often called Mantel–Haenszel tests, but it is easy to extend such tests to ordinal items and all types of covariates.

Mantel–Haenszel techniques are widely used in all IRT models motivated by the intuition that since the IRT model assumes that the item and the covariate are conditionally independent given the unobserved θ, then it must surely also be true when the total score is used as a proxy for the latent variable. Unfortunately, such an intuition is once again wrong. The only known IRT model where conditional independence between an item and a covariate given the latent θ implies that the item and the covariate are conditionally independent given the total score R is the Rasch model. The hypothesis may be almost true (in some unspecified sense) under other models, but is nevertheless false and it will result in erroneous evidence of DIF in large sample studies where items fit other types of IRT models. In other words, if you have a very large sample and if you do not find any evidence against conditional independence after stratification by the total score, then you have every reason to believe that your items follow something similar to a Rasch model whether or not that was what you aimed for to begin with.

EXAMPLE 5.7.– To illustrate what goes on during such an analysis, we show the relationship between item DHP36 and gender in two score strata (Table 5.7).

Score		0	1	2	3	Total	
6	Male	1	1	4	3	9	
	%	11.1	11.1	44.4	33.3	100.0	
	Female	2	2	4	0	8	
	%	25.0	25.0	50.0	0.0	100.0	
	Total	3	3	8	3	17	$\chi^2 = 3.6$, df $= 3$, $p = 0.305$
	%	17.6	17.6	47.1	17.6	100.0	$\gamma = -0.62$, $p = 0.023$
7	Male	1	4	6	1	12	
	%	8.3	33.3	50.0	8.3	100.0	
	Female	4	3	6	1	14	
	%	28.6	21.4	42.9	7.1	100.0	
	Total	5	7	12	2	26	$\chi^2 = 1.8$, df $= 3$, $p = 0.615$
	%	19.2	26.9	46.2	7.7	100.0	$\gamma = -0.23$, $p = 0.220$

Table 5.7. *Association between DHP and gender in two strata with total score 6 and 7, respectively*

In Table 5.7, the strength of the conditional association between item DHP36 and gender is measured by Goodman and Kruskal's γ because this measure is appropriate for ordinal categorical data. In both tables, males have higher item scores than females, but the difference is only significant in the stratum containing respondents with score 7. To summarize the results over all strata, we use the partial γ coefficient [DAV 67] and use Monte Carlo estimates of exact conditional p-values [KRE 87] to avoid problems with inappropriate asymptotic results in large sparse contingency tables.

The partial γ is -0.32 with $p = 0.01$, confirming the results of Kelderman's CLR test. Gender is a source of DIF for item DHP36.

5.8. Discussion and recommendations

There are several topics that we need to discuss regarding this chapter, including technical problems relating to the choice of item fit statistics and the more fundamental problems concerning what to do, when item fit statistics disclose evidence against the item-model fit.

5.8.1. *Technical issues*

Let us take the technical problem first. A lot of fit statistics for Rasch models have been suggested and several statistics have been implemented in available software for Rasch models. Most, if not all, suggestions are based on sound ideas and on unquestionable knowledge about Rasch measurement (as opposed to Rasch inference), but the implementation of these methods in many cases is limited and based on the lack of understanding of statistical inference, to a degree where we expect that many published results based on these implementations are erroneous. The best advice we can give to the users of Rasch models is that they should search for implementations of methods based on the principles of conditional inference because conditional inference guarantees that results are consistent without bias and less errorneous in large sample studies. Other methods may – it has to be admitted – work almost as well as conditional methods in small sample studies but they are guaranteed to have bias and to inflate the type I error rates as sample sizes increase because implementation of fit statistics that are not based on conditional inference tests hypotheses that at best are almost true, but always false. This, at least, is what statistical theory tells us. If you doubt this and want to see what happens when you analyze data using your favorite Rasch program, we suggest that you simulate data from Rasch models and take special care to see what happens when sample sizes increase. If the program systematically rejects the model as sample sizes increase, you can be sure that your program has a problem.

We have already referred to Glas and Verhelst [GLA 95a, GLA 95b] for an overview of item fit statistics and additional references. We would also like to point out that item fit statistics developed for other IRT models ([GLA 99] is a particular useful reference) are also useful for the assessment of item fit to Rasch models because the Rasch model is a special case of these models.

If we find, for instance, that an item does not agree with the non-parametric IRT model, it follows automatically that it cannot agree with the Rasch model. Christensen

and Kreiner [CHR 07] have pursued this point and have found that the power of the Loevinger scalability H coefficient derived from non-parametric IRT models compares favorably to the power of fit statistics developed within the theory of Rasch models.

Given the large number of available item fit statistics, it is common for new users of Rasch models to be confused. We would, of course, always search for the most powerful statistics, but there are few studies comparing the power of item fit statistics in a systematic and comprehensive way. For what it is worth, we can inform the reader that we have performed a number of (unpublished) Monte Carlo studies without finding indisputable evidence that some of the fit statistics are more powerful than others to a degree that makes a difference in practice. So feel free to choose any fit statistic that has captured your attention. However, that does not mean that your choice of item fit statistic should be indiscriminate. It is your responsibility to ensure that the implementation of these methods is up to standard. You will find, when you look into these matters, that many of these methods were developed and implemented in the 1970s and early 1980s based on technology available at that time, but the developments in statistical inference, in general, and psychometric methods, in particular, since then have been ignored. This is regrettable but also common, and there are no reasons to complain to those who at some point have invested time and money in developing these programs. The responsibility solely lies with the user. It is the responsibility of the user to check that the methods have been implemented in ways that guarantee against other errors than those we cannot avoid during statistical analyses because of random variation in data.

5.8.2. *What to do when items do not agree with the Rasch model*

It is important to understand that the item fit statistics targeting the item-model fit and, in particular, the item fits statistics checking the response homogeneity do not question the fundamental assumptions of unidimensionality, local independence and no DIF. The item fit statistics are only concerned about the choice of the probability functions and not about anything else. Evidence against item-model fit may turn up – or rather will turn up – if the fundamental assumptions of IRT and Rasch models are violated. You are therefore required to look into all these issues before you decide to either eliminate an apparently bad item or choose a different probability function – a different IRT model – to obtain a better fit to data. You should be aware, for instance, that local dependence will have the effect that it appears as if items have different item discriminations and that local dependence itself may be caused by multidimensionality and/or DIF relative to observed and unobserved covariates. Evidence of item misfit is therefore often only the first step of a long (and sometimes tedious) journey to find out what is really wrong and what to do about it. Needless to say – but let us say it,

nevertheless, to ensure that we do not forget it – that you are also required to take a second look both at the subject matter arguments that lead to the development of your items and the contents and formulations of questions and response categories. You will find – as we have found many times – that items are rejected not because of the content of the items but because of inept, unprofessional and careless item writing. Local response dependence and DIF are typically caused by these reasons, and when this happens, there are only three options: you eliminate the items, rewrite the items or select an IRT model that adjusts for the local dependence and the DIF. In other cases, the lack of item-model fit challenges the subject matter arguments leading to your items. In our experience, this happens less often than because of bad item writing, but it does happen, in particular, in connection with what we could refer to as *ad hoc* scales. When this happens, you must rethink the whole theoretical foundation underlying the items. Exploratory methods searching for multidimensional models providing a better fit to data may be of help here, but they cannot stand alone without serious subject matter considerations. Item analyses by Rasch models and other IRT models are confirmatory analyses, which assume that you know what you are doing but acknowledge that you may be wrong and therefore want to check that you have not made any errors. Turning to exploratory methods when data-model fit fails is the same as admitting that you did not know anything and, therefore, need the computer to sort things out, and that, of course, cannot be true. You do know something, but the lack of fit tells you that you have made some errors. The best solution to that problem is to think again.

5.9. Bibliography

[AND 04] ANDRICH D., "Controversy and the Rasch model: a characteristic of incompatible paradigms?", *Medical Care*, vol. 42, pp. 7–16, 2004.

[BEN 95] BENJAMINI Y., HOCHBERG Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing", *Journal Royal Statistical Society, Series B*, vol. 27, pp. 313–324, 1995.

[CHR 07] CHRISTENSEN K.B., KREINER S., "A Monte Carlo test approach to unidimensionality testing in polytomous Rasch models", *Applied Psychological Measurement*, vol. 31, pp. 20–30, 2007.

[DAV 67] DAVIS J.A., "A partial coefficient for Goodman and Kruskall's gamma", *Journal of the American Statistical Association*, vol. 69, pp. 37–46, 1967.

[GLA 95a] GLAS C.A.W., VERHELST N.D., "Testing the Rasch model", in FISCHER G.H., MOLENAAR I.W. (eds), *Rasch models: Foundations, Recent Developments, and Applications*, Springer-Verlag, New York, NY, pp. 69–95, 1995.

[GLA 95b] GLAS C.A.W., VERHELST N.D. "Tests of fit for polytomous Rasch models", in FISCHER G.H., MOLENAAR I.W. (eds), *Rasch models: Foundations, Recent Developments, and Applications*, Springer-Verlag, New York, NY, pp. 325–352, 1995.

[GLA 99] GLAS C.A.W., "Modification indices for the 2-pl and the nominal response model", *Psychometrika*, vol. 64, pp. 273–294, 1999.

[GOO 54] GOODMAN L.A., KRUSKAL W.H., "Measures of association for cross classifications", *Journal of the American Statistical Association*, vol. 49, pp. 732–764, 1954.

[KEL 84] KELDERMAN H., "Loglinear Rasch model tests", *Psychometrika*, vol. 49, pp. 223–245, 1984.

[KRE 87] KREINER S., "Analysis of multidimensional contingency tables by exact conditional tests: techniques and strategies", *Scandinavian Journal of Statistics*, vol. 14, pp. 97–112, 1987.

[KRE 11] KREINER S., "A note on item-restscore association in Rasch models", *Applied Psychological Measurement*, vol. 35, pp. 557–561, 2011.

[MOL 83] MOLENAAR I.W., "Some improved diagnostics for failure of the Rasch model", *Psychometrika*, vol. 48, pp. 49–72, 1983.

[SMI 04] SMITH R.M., "Fit analysis in latent trait measurement models", in SMITH E.V., SMITH R.M. (eds) *Introduction to Rasch measurement*, JAM Press, Maple Grove, Minnesota, pp. 73–92, 2004.

[VAN 82] VAN DEN WOLLENBERG A.L., "Two new test statistics for the Rasch model", *Psychometrika*, vol. 47, pp. 123–139, 1982.