# Chapter 16
# Fit of Responses to the Model III—Differential Item Functioning

*Statistics Review 12: Analysis of variance (ANOVA)*

The first section of this chapter describes how to visualize DIF from the ICC; the next section deals with how to confirm DIF statistically; finally, the concepts of artificial DIF and resolving items with DIF are introduced.

Differential item functioning (DIF) occurs when items do not function in the same way for different groups of people, who otherwise have the same value on the trait. DIF refers to items having different relative difficulty for groups and therefore violating invariance, and has been referred to as *bias*. It does not refer directly to one group of people having a greater score than another group on the item. In developing a new measure, whether it is an achievement test or a questionnaire, it is important to investigate whether the items have different meanings for different groups (e.g. male/female, employed/unemployed, married/not married). If valid quantitative comparisons are to be made among groups, the item parameters need to be invariant across the groups to be compared. This *measurement* requirement of invariance seems to have been first articulated by Thurstone:

> If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help construct it. This may turn out to be a severe test, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale (Thurstone, 1928, p. 547 cited in Andrich & Hagquist, 2004).

The property of invariance quoted from Thurstone above implies a requirement of the data. Any model can be applied to different subgroups in order to investigate whether or not there is invariance amongst the parameters estimated. The main advantage of the Rasch model in the study of invariance is that it has this property built into its own structure. Its general form (Rasch, 1961) was developed from the requirements that

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.

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_16

> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch, 1961, p. 322 cited in Andrich & Hagquist, 2004).

In principle, two different approaches can be taken to identify DIF. One approach is to estimate a single set of parameters for each item and then study the residuals identified by the different groups. For example, if the groups were boys and girls, one would analyse boys' and girls' responses in the same analysis and compare a mean residual for boys with a mean residual for girls. Another approach is to estimate parameters in different groups and then compare the estimates. Both of these approaches are described in detail in Andrich and Hagquist (2004). In this chapter, we look at the first approach and describe this approach using the example from Andrich and Hagquist.

The example involves survey data from a study collected in 1998 among Year 9 students in Sweden. The data collection involved a questionnaire including eight items intended to be a measure of well-being and perceived health. The questions were '*During this school year, have you…*' felt that you have had difficulty in concentrating? felt that you have had difficulty in sleeping? suffered from headaches? suffered from stomach aches? felt tense? had little appetite? felt low? felt giddy? The response categories for all the items were *never*, *seldom*, *sometimes*, *often* and *always*. The total number of persons used in the analysis was 654, with 301 boys and 353 girls.

It was important to assess if the survey items functioned the same way for boys and girls, that is, that they are not biased towards one group. DIF can be visualized or detected graphically by means of the item characteristic curve (ICC). It can also be confirmed statistically.

## Identifying DIF Graphically

There is a vast literature on DIF. From the perspective of detecting DIF in this book, we focus on the item characteristic curve (ICC). Thus a single ICC is estimated for all persons irrespective of group membership, and then the observed means of responses in class intervals across the continuum should be close to the ICC. In the case of dichotomous responses, the observed means are the proportions of positive responses, and the expected value is simply the probability of a positive response.

Within the tradition of modern test theory, the fundamental idea of *no* DIF among groups is that for the same values of the trait, the expected value of a member from any group of individuals is identical. The expected values are displayed in an item's ICC. From this perspective of an invariant ICC, there are in principle *three basic kinds of DIF*:

(i) the *locations* of the curves are *different* in the different groups but their *slopes* are the *same*. DIF with parallel slopes is referred to as *uniform DIF*.

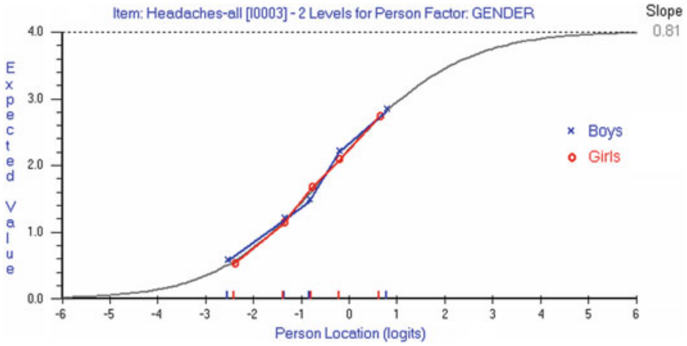

**Fig. 16.1** Graphical comparison between means of boys and girls in 5 class intervals for item 3 showing no systematic difference between genders

(ii) the *locations* are the *same* but their *slopes* are *different*. DIF with non-parallel slopes is referred to as *non-uniform DIF*.

(iii) both their *slopes* and their *locations* are *different.* This DIF is also called *non-uniform DIF*.

To check graphically if an item has DIF between boys and girls we can look at the ICC for the item. The observed means in class intervals are displayed separately for boys and girls. The ICC for item 3 (headaches) is shown in Fig. 16.1. The graph shows that the observed means of the boys and girls in the class intervals are both close to each other and close to the expected values. This evidence indicates that the item fits the model and that there is no DIF.

The graph for the two groups is very different in Fig. 16.2, which shows the ICC for item 7 (felt low). For the same class interval, that is mean person location, girls have systematically higher observed means than boys. The observed means of the girls are greater than expected and of the boys less than expected. Item 7 shows *uniform* DIF. If the slopes were not parallel and crossed, they would have shown non-uniform DIF.

## Identifying DIF Statistically Using ANOVA of Residuals

Whilst the graphical display gives a visual orientation to the data, DIF can be confirmed statistically through an analysis of the residuals. The standardised residual of each person $n$ to each item i is given by

$$z_{ni} = \frac{x_{ni} - E[x_{ni}]}{\sqrt{V[x_{ni}]}}, \tag{16.1}$$

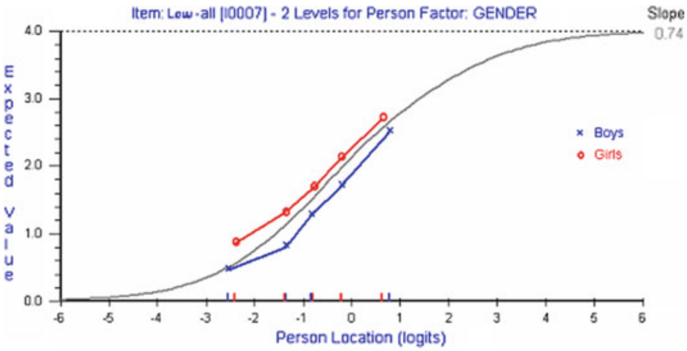

**Fig. 16.2** Graphical comparison between means of boys and girls in 5 class intervals for item 7 showing a DIF effect between genders

where $E[x_{ni}]$ is the expected value given person $n$'s and item $i$'s parameter estimates, and $V[x_{ni}]$ is the variance.

For the purpose of the detailed analysis, each person is further identified by the gender group $g$ and by the class interval $c$. This gives the residuals

$$z_{n_{cg}i} = \frac{x_{n_{cg}i} - E[x_{n_{cg}i}]}{\sqrt{V[x_{n_{cg}i}]}}. \tag{16.2}$$

These residuals are analysed according to standard analysis of variance (ANOVA). ANOVA is a statistical procedure used to determine whether there is a significant difference between the means of two or more groups. In the case of identifying DIF, ANOVA is used to determine whether there is a significant difference among the mean residuals for the groups of interest, in this case boys and girls. The question about whether means are different is answered by analysing the variation among the means. In an analysis of variance, F-ratios are constructed. An F-ratio is a ratio of the estimated variance of residuals among groups and the estimated variance of residuals within groups. Under the assumption that the means come from a single random set of residuals from within the groups, then the theoretical value of this ratio is 1.0. If the F-ratio is greater than 1.0, this could indicate that there is a real difference among the group means. How much greater than 1 does it need to be before we can say the group means are significantly different?

Because we are working with estimates of variances in ANOVA we cannot say with 100% certainty whether an observed difference is real. It may just be a chance difference, a peculiarity of the particular sample of residuals. In the ANOVA output below, both an F-ratio and a probability are given. If the probability is less than a certain chosen criterion one can conclude that the difference between the means is statistically significant. That is, the F-ratio of this magnitude would occur by chance less often than indicated by the probability. If the probability is greater than the chosen criterion one can conclude that the difference is not statistically significant.

Values of 0.01 or 0.05 are the criteria chosen most often. If the difference between the means of boys and girls are significant we say there is a *main effect* of gender. Please refer to *Statistics Review 12* for an explanation of the concepts underlying analysis of variance.

Table 16.1 provides a summary of the analysis of variance of the residuals with the F-ratio and its significance for each of the eight items for (i) the uniform DIF gender effect, (ii) the non-uniform DIF interaction effect, and (iii) the class interval effect.

Returning to the graphs in Figs. 16.1 and 16.2, it would be expected that there is not a significant main effect of gender for item 3 and that there is a significant main effect of gender for item 7. For item 3, we are interested in the gender F-ratio of 0.024 which is not statistically significant according to the chosen criterion of 0.01 ($p = 0.871$). There is not a main effect of gender. For item 7 the main effect of gender is statistically significant. The gender F-ratio is 44.543 and the probability is 0.000. This confirms the uniform DIF we identified graphically in Fig. 16.2.

In the first section, we noted that there are in principle three basic kinds of DIF. An ANOVA main effect confirms uniform DIF, that is, the locations of the curves are different in the different groups but their slopes are parallel. To determine whether the slopes are not parallel for the different groups, i.e. to confirm non-uniform DIF, we need to look at another type of effect in ANOVA called an *interaction effect*. An interaction effect occurs when the residuals are different for different groups depending on the class intervals.

**Table 16.1** Analysis of variance of residuals for the test of DIF between genders taken from Andrich and Hagquist (2004)

| Items | | ANOVA | | | | | |
|---|---|---|---|---|---|---|---|
| | | Gender | | Gender by class interval | | Class interval | |
| | | F (df: 1, 639) | $p <$ | F (df: 4, 639) | $p <$ | F (df: 4, 639) | $p <$ |
| **1** | Concentrating? (B > G) | 11.744 | **0.001** | −0.671 | N/Sig | 2.252 | 0.061 |
| **2** | Sleeping? (B > G) | 8.882 | **0.003** | 1.155 | 0.329 | 0.521 | 0.723 |
| 3 | Headaches? | 0.024 | 0.871 | 0.569 | 0.688 | 0.186 | 0.944 |
| **4** | Stomach aches? (G > B) | 19.362 | **0.000** | 1.418 | 0.225 | 1.825 | 0.121 |
| 5 | Tense? | 0.003 | 0.958 | 0.526 | 0.720 | 2.154 | 0.072 |
| 6 | Appetite? | 1.316 | 0.250 | 0.697 | 0.597 | 0.831 | 0.508 |
| **7** | Low? (G > B) | 44.543 | **0.000** | 0.597 | 0.668 | 0.951 | 0.565 |
| 8 | Giddy? | 5.800 | 0.015 | 1.537 | 0.189 | 0.407 | 0.806 |

Number of class intervals = 5; $p < 0.01$ taken as significant

The gender by class interval interaction effect indicates whether the discrimination or slope of the item for the two genders is different. For item 3 the gender by class interval F-ratio is 0.569 which is not statistically significant ($p = 0.688$). For item 7 the gender by class interval F-ratio is 0.597 which is also not statistically significant ($p = 0.668$). This confirms no non-uniform DIF for items 3 and 7 in Figs. 16.1 and 16.2.

In addition to confirming the graphical interpretations from Figs. 16.1 and 16.2, Table 16.1 shows that for item 4 (*Suffered from stomach aches*) there is also gender DIF. In particular, there is a greater prevalence of ailment in girls (for the same overall location). Item 1 (*Difficulty in concentrating*) and item 2 (*Difficulty in sleeping*) show marginal greater prevalence of ailments in boys than girls (again, for the same overall location). The above evidence was used to conclude that four items, items 1, 2, 4 and 7 show DIF, which is primarily uniform. This, of course, implies that the remaining four items were taken not to show DIF. Item 8 (*Felt giddy*) did not show either uniform or non-uniform DIF. In that analysis, the DIF criterion was set at the 0.01 level and so the main effect of gender was taken as not significant.

In summary, two ANOVA effects are relevant to consider for DIF. The first is whether there is a main group effect, and the second whether there is a group by class interval effect. The former indicates whether or not the mean of the size of the residuals of the two groups on the average is different. The latter indicates whether the discrimination or slope of the item for the two groups is different.

For completeness, we can consider the class interval effect. This gives analogous information to the $\chi^2$ test across intervals. That is, it checks whether, irrespective of groups, the mean residuals are statistically equivalent among class intervals. If these are significantly different, then that implies that the actual means are not close to the theoretical curve.

In our discussion on DIF, we have focused on just two groups. We consider them of equal status. In some work on DIF, the perspective is that there is a standard or main group and that there is a subgroup, sometimes referred to as a *focal* group, which might have items which are biased against it. However, we do not take that perspective here. Indeed, we suggest that unless there is some special reason, the sample sizes of the two groups should be as close as possible to the same. This is because if the sample sizes are different, and there is DIF, then the estimates will be weighted by the estimates that would be present for the group with the larger sample size.

In the previous section, we discussed detecting DIF. In the next section, we discuss ways of studying DIF more closely, as well as the concepts of artificial DIF and resolving items. Sometimes the term splitting is used, but we use resolving to convey the sense of showing the constituent parts of the DIF.

## Artificial DIF

Andrich and Hagquist (2012) introduce the concept of artificial DIF. The basic issue is that in forming class intervals, the known values of persons on the continuum are not known—only estimates are known. When class intervals are formed using estimates, which is effectively the same total score, then if some item has a higher value for a class interval, and the total score for the class interval across items is fixed, then other items must have some lower values. Thus, the artificial DIF favours the group opposite to that of the real DIF. Whether or not it becomes noticeable depends on other features of the data.

For the rest of the illustrations of this chapter, a set of data was simulated to represent 1000 boys and 1000 girls. There were 8 items, each with 4 categories. All items have the same location and the same thresholds for boys and girls, except item 3 was simulated to have a location value with a difference between boys and girls of 0.71 logits favouring boys.

Table 16.2 shows that there is no non-uniform DIF, and there is no misfit across the continuum as evidenced by the class interval fit statistics. However, two items show misfit due to gender. One is item 3 which is expected because of the simulation. However, item 4 also shows DIF. This item shows artificial DIF.

A RUMM2030 analysis of the real data shown in Table 16.1, which is explained in more detail below, also shows this effect.

In this example, artificial DIF manifested itself in item 4 most noticeably, though in theory there is a small artificial effect in all items. The reason it showed statistical significance in item 4, and not other items, is that item 4 must have, by chance, had some DIF favouring girls, and with the extra effect of artificial DIF, it showed up.

The magnitude of real DIF and incidents of artificial DIF, on item parameter estimates can be quantified. This is done by resolving the items into group specific items. To quantify DIF, items showing DIF must be resolved sequentially, and in particular, if there is more than one item that shows DIF, then the one to deal with first is the one which has the highest Mean Square. Thus if item 3 is resolved, then

**Table 16.2** DIF summary with items 3 and 4 significant at this level for gender effect

| Item | Class interval | | | | Gender | | | | Class interval by gender | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MS | F | DF | Prob | MS | F | DF | Prob | MS | F | DF | Prob |
| I0001 | 2.445 | 3.129 | 9 | 0.001 | 0.054 | 0.069 | 1 | 0.793 | 1.306 | 1.671 | 9 | 0.091 |
| I0002 | 1.173 | 1.327 | 9 | 0.217 | 0.108 | 0.121 | 1 | 0.728 | 1.558 | 1.762 | 9 | 0.071 |
| I0003 | 1.219 | 1.521 | 9 | 0.135 | **79.599** | **99.301** | **1** | **0.000** | 1.328 | 1.656 | 9 | 0.094 |
| I0004 | 0.510 | 0.563 | 9 | 0.828 | **18.739** | **20.711** | **1** | **0.000** | 0.916 | 1.012 | 9 | 0.428 |
| I0005 | 1.369 | 1.544 | 9 | 0.127 | 0.096 | 0.108 | 1 | 0.742 | 0.762 | 0.859 | 9 | 0.562 |
| I0006 | 1.537 | 1.813 | 9 | 0.061 | 2.888 | 3.407 | 1 | 0.065 | 1.436 | 1.694 | 9 | 0.085 |
| I0007 | 2.390 | 2.934 | 9 | 0.002 | 0.507 | 0.622 | 1 | 0.430 | 0.842 | 1.034 | 9 | 0.410 |
| I0008 | 1.753 | 2.120 | 9 | 0.025 | 3.374 | 4.081 | 1 | 0.044 | 0.428 | 0.517 | 9 | 0.863 |

the rest of the items should fit. However, if there is another item then that shows DIF, it would be resolved, and so on.

## Resolving Items

Resolving item 3 in this example, means creating two new items, one responded to only by boys and the other only by girls. The resolved item does not show DIF in the ANOVA because boys and girls now have distinct items. In RUMM2030 it is simply referred to as *split*.

The resolution of an item creates missing responses in some cells of the data matrix. If it is the only item with real DIF, and it generated artificial DIF in any other items, then when the item is resolved the artificial DIF effect will not be present in an analysis of the modified matrix.

Table 16.3 shows the ANOVA of residuals after item 3 has been resolved. At the same level of statistical significance as in Table 16.2, no item shows misfit. It is evident that item 4 has a marginal misfit, and indeed it shows marginally higher scores for girls. The artificial DIF put it over the significance limit.

Figure 16.3 shows the ICCs for boys and girls for item 3. It also shows the observed means in the class intervals, which are close to their respective curves. Recall that the overall item location difference that was simulated was 0.71. From the location estimates in Fig. 16.3 the estimated difference is $0.44 - (-0.31) = 0.75$. This is a very good estimate of the effect that was simulated. The slope estimates reflect the threshold estimates, and relative to the mean of the thresholds they were identical with a difference of only 0.07.

Thus resolving the items in this way gives an excellent, theoretically sound way of estimating the effect of DIF in terms of the parameters of the items.

**Table 16.3** DIF summary after item 3 is resolved: no significance at this level

| Item | Class interval | | | | Gender | | | | Class interval by gender | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MS | F | DF | Prob | MS | F | DF | Prob | MS | F | DF | Prob |
| I0001 | 2.3982 | 3.0344 | 9 | 0.0013 | 1.3065 | 1.6531 | 1 | 0.1987 | 1.2281 | 1.5538 | 9 | 0.1238 |
| I0002 | 0.9771 | 1.0994 | 9 | 0.3597 | 0.7221 | 0.8125 | 1 | 0.3675 | 1.6876 | 1.8989 | 9 | 0.0480 |
| I0004 | 0.8938 | 0.9837 | 9 | 0.4513 | 9.1339 | 10.0524 | 1 | 0.0016 | 0.7556 | 0.8316 | 9 | 0.5870 |
| I0005 | 1.1879 | 1.3305 | 9 | 0.2156 | 0.9769 | 1.0941 | 1 | 0.2957 | 0.8672 | 0.9713 | 9 | 0.4618 |
| I0006 | 1.6465 | 1.9283 | 9 | 0.0441 | 0.1625 | 0.1903 | 1 | 0.6627 | 1.0606 | 1.2421 | 9 | 0.2645 |
| I0007 | 2.1511 | 2.6181 | 9 | 0.0052 | 0.4434 | 0.5397 | 1 | 0.4627 | 0.9264 | 1.1275 | 9 | 0.3394 |
| I0008 | 1.5762 | 1.8877 | 9 | 0.0496 | 0.1439 | 0.1723 | 1 | 0.6782 | 0.4738 | 0.5674 | 9 | 0.8247 |
| Girls | 1.1452 | 1.4603 | 9 | 0.1581 | 0.0000 | 0.0000 | 0 | 1.0000 | 0.0000 | 0.0000 | 0 | 1.0000 |
| Boys | 0.9829 | 1.1782 | 9 | 0.3053 | 0.0000 | 0.0000 | 0 | 1.0000 | 0.0000 | 0.0000 | 0 | 1.0000 |

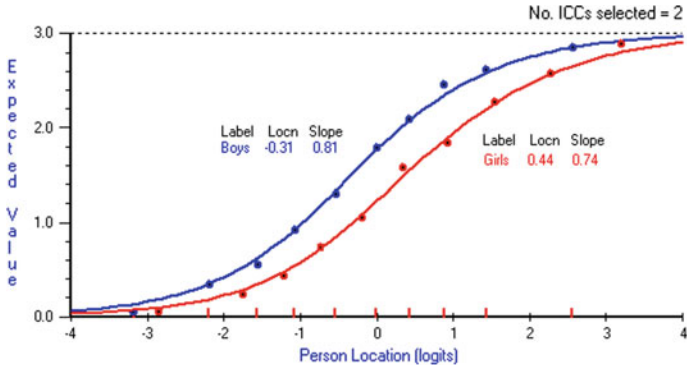

**Fig. 16.3** Resolved item 3 for boys and girls

## Exercises

*Exercise 2: Basic analysis of dichotomous and polytomous responses* in Appendix C.

*Exercise 5: Analysis of data with differential item functioning* in Appendix C.

## References

Andrich, D. & Hagquist, C. (2004). *Detection of differential item functioning using analysis of variance*. Paper presented at the Second International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch Models.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioural Statistics, 37*(3), 387–416.

## Further Reading

Andrich, D., & Hagquist, C. (2015). Real and artificial differential item functioning in polytomous items. *Educational and Psychological Measurement, 75*(2), 185–207.

Broderson, J., Meads, D., Kreiner, S., Thorsen, H., Doward, L., & McKenna, S. (2007). Methodological aspects of differential item functioning in the Rasch model. *Journal of Medical Economics, 10,* 309–324.

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences, 36,* 955–968.

Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF—A study based on simulated polytomous data. *Psychological Test and Assessment Modelling, 57,* 342–376.

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes, 15*(181), 1–8.

Looveer, J., & Mulligan, J. (2009). The efficacy of link items in the construction of a numeracy achievement scale—From kindergarten to year 6. *Journal of Applied Measurement, 10*(3), 247–265.