Difficulty plots: DIF analysis without the partial credit model

Elizabeth Harrison

Introduction

From summer 2017, the preferred method for vertical scaling of tiered exams will be chained equipercentile equating (JCQ, 2015; Kolen & Brenan, 2004). AQA has previously used the partial credit model (PCM) to perform vertical scaling (see Kolen & Brennan, 2004; Harrison & Pointer, 2016). For either method, differential item functioning (DIF) should be assessed among the common items, ideally at the time of the award (JCQ, 2015). Any items that perform very differently on the foundation and the higher tier might need to be dropped from the common item pool to ensure a fair link between the two tiers. DIF is fairly straightforward to assess when using the PCM; this report illustrates a simple method, using difficulty plots, to assess DIF without fitting the PCM. The two approaches are compared over 12 tiered GCSE papers from the summer 2016 series.

Illustration comparing PCM DIF and difficulty plots

Table 1 shows some summary statistics for the common items for GCSE Gardening.[1] The item-level parameters β_F and β_H are from the fitted PCMs (after calibration[2]) and describe the difficulty of each item. The facility is the mean mark divided by the tariff; it is also a measure of the item difficulty.

Table 1 Common item statistics for GCSE Gardening

F item	H item	tariff	β_F	β_H	F facility	H facility	Orthogonal residual	PCM DIF (β_F-β_H)
8a	1a	1	0.32	0.75	0.28	0.49	-0.020	-0.43
8bi	1bi	1	-0.17	0.80	0.39	0.48	-0.103	-0.97
8bii	1bii	1	0.31	0.48	0.30	0.56	0.019	-0.17
8biii	1biii	1	5.20	3.19	0.01	0.10	-0.106	2.01
8biv	1biv	2	1.63	1.13	0.10	0.42	0.051	0.49
8c	1c	3	0.61	0.98	0.36	0.53	-0.043	-0.36
9a	2a	1	2.29	1.05	0.06	0.43	0.092	1.24
9bi	2bi	1	-0.24	0.21	0.40	0.61	-0.019	-0.44
9bii	2bii	1	1.88	1.25	0.08	0.38	0.041	0.63
9c	2c	1	-1.00	-0.97	0.58	0.81	0.002	-0.03
9d	2d	6	0.28	0.57	0.20	0.56	0.085	-0.29

[1] The data is from a real GCSE qualification but a fictional subject name has been used.

[2] Calibration is the process of aligning the scales of the PCMs fitted to the foundation and higher tiers.

There is a strong, negative correlation between the β estimates and the item facilities over the 12 papers analysed here: ranging from -0.85 to -0.99 (high β values are equivalent to low item facilities). The final two columns of Table 1 are measures of DIF and will be explained later in this report.

The relative difficulty of each common item on each of the two tiers can be summarised graphically using a difficulty plot (Livingstone, 2004, p. 32). Figure 1 shows the relative difficulties of the common items from GCSE Gardening. The line shown is the orthogonal regression line.[3] Orthogonal regression assumes that there is error in the measurement in the observations on both the *x*- and *y*-axis. It is also symmetric; hence the same items would be flagged if *y* were regressed on *x* or *x* regressed on *y*. It assumes that the error variances are equal. (Simple linear regression assumes there is only error in observations on the *y*-axis). A residual can then be defined as the orthogonal distance to the regression line; this distance is given in Table 1. Figure 1 shows item 8biii as an outlier, together with a few others (8bi, 9a and 9d) that might need investigating.

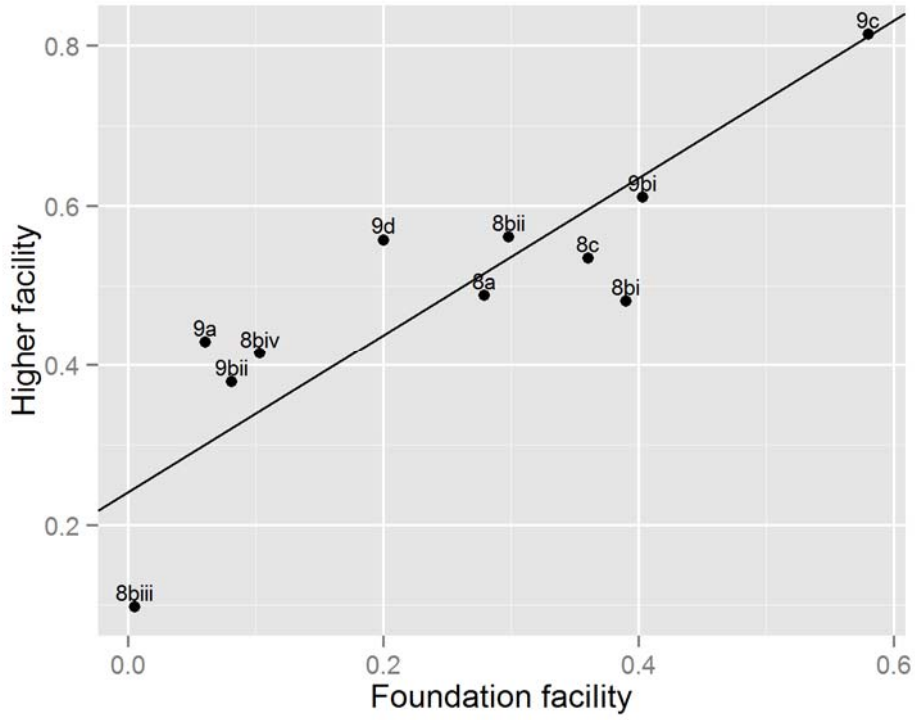

Figure 1 Difficulty plot for GCSE Gardening

Figure 2 shows the PCM difficulty (β) estimates at threshold level for the common items of the GCSE Gardening papers. Each threshold for an item is labelled; for example, 8c, a three-mark item, is shown as 8c.1, 8c.2, 8c.3. The thresholds describe the difficulty associated with acquiring each individual extra mark in a polytomous item (Masters, 1982). This level of detail is lost when using difficulty plots as we are only looking at average performance rather than how performance changes across the mark range. A rule of thumb for flagging problematic PCM DIF is to investigate items with an absolute DIF of 0.5 or more at any threshold level. Figure 2 shows that item 8biii has very different difficulty (β) estimates for each tier – this is evidence of DIF.

[3] Fitted using the R package *mcr*

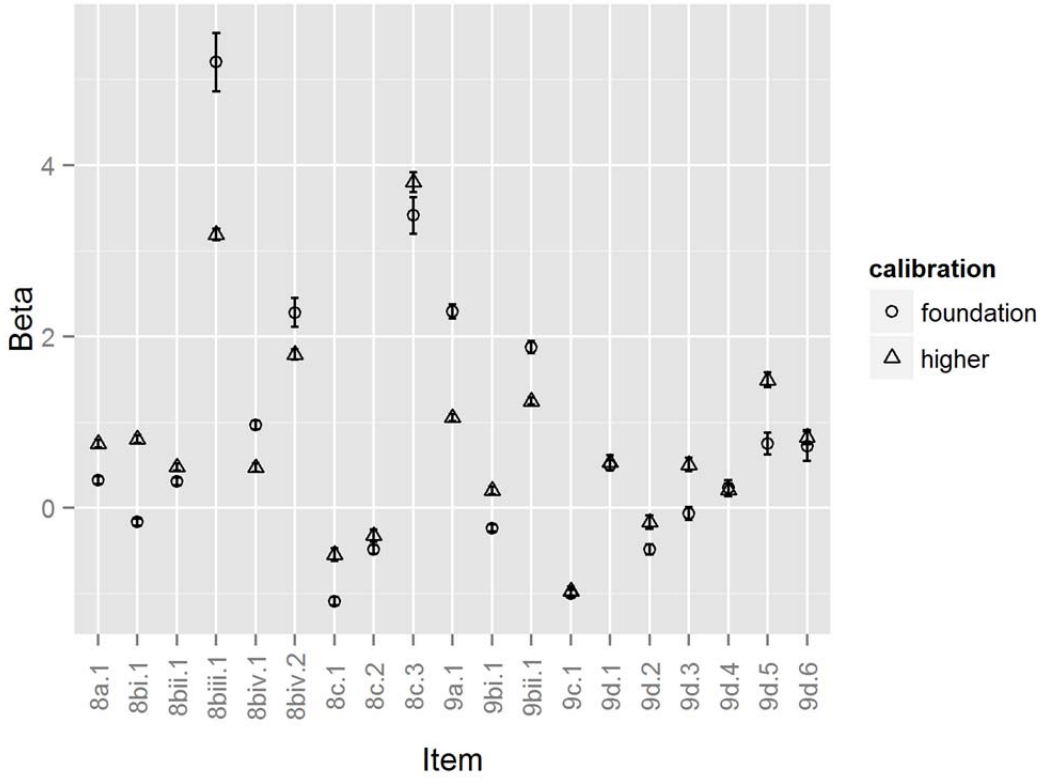

Figure 2 **Comparison of calibrated β estimates for GCSE Gardening, with error bars**

The differences (DIF) between the β estimates are summarised in Table 1, where they are shown at item level for more direct comparison with orthogonal residuals (i.e. for a polytomous item, the average DIF over all the item's thresholds is given).

Assessing DIF is largely a matter of judgement, and the appropriateness of removing an item from the common item pool would need to be confirmed by looking at the exam paper and consulting subject experts. DIF could be caused by a common item being closely linked to a unique item on one of the tiered papers (e.g. a non-common item may give a strong hint as to how to answer a common item), or by being unrepresentative of the difficulty at the overlapping grades, particularly if it is too hard for the foundation tier candidates.

Ideally, very few items should show DIF; here, GCSE Gardening shows more DIF than is usually seen. Table 1 and Figure 2 show four items with high PCM DIF (greater than 0.5): 8bi, 8biii, 9a and 9bii. If an absolute residual of 0.1 (10%) is used as the cut-off to flag items with DIF on the difficulty plot, then items 8bi and 8biii are again flagged. Item 8biii appears particularly difficult for both tiers and might not be representative of the desired common item difficulty. Item 9a does have a high residual of 0.092; however, item 9bii does not. Conversely, item 9d has a fairly high residual of 0.085, but the PCM DIF (0.29) does not reflect this. This item is a six-mark item; it does show DIF at some thresholds (see Figure 2) but only a little once averaged over all the thresholds.

The two approaches do not entirely agree, but this is not surprising. The PCM DIF is sensitive to the adequacy of the PCM fit and the quality of the link defined between the tiers. The regression line, though much simpler to create, is sensitive to items whose facility values cause leverage (i.e. the position of a point in the difficulty plot may have a strong influence on the final estimated regression equation).

An aside on PCM DIF and ability

It should be noted that where PCM DIF exists, its impact on individual students is dependent on their individual ability, relative to the difficulty of the item. When an item is particularly easy or difficult, relative to a student's ability, DIF will have little impact. However, when a candidate's ability is close to the difficulty of the item, the likelihood of a correct answer is very different depending on the tier. This is illustrated in Figure 3, which shows the increased chances of success for higher tier candidates over foundation tier candidates for a dichotomous item, over a range of abilities and for two values of DIF that favour the higher tier (0.5 and 1.0).[4] Therefore, large, positive DIF in a common item that was targeted at candidates of grade C ability might be detrimental to the foundation tier, and adversely affect the vertical scaling. However, DIF in a very difficult item, such as item 8biii, might not have much impact on the vertical scaling analysis.

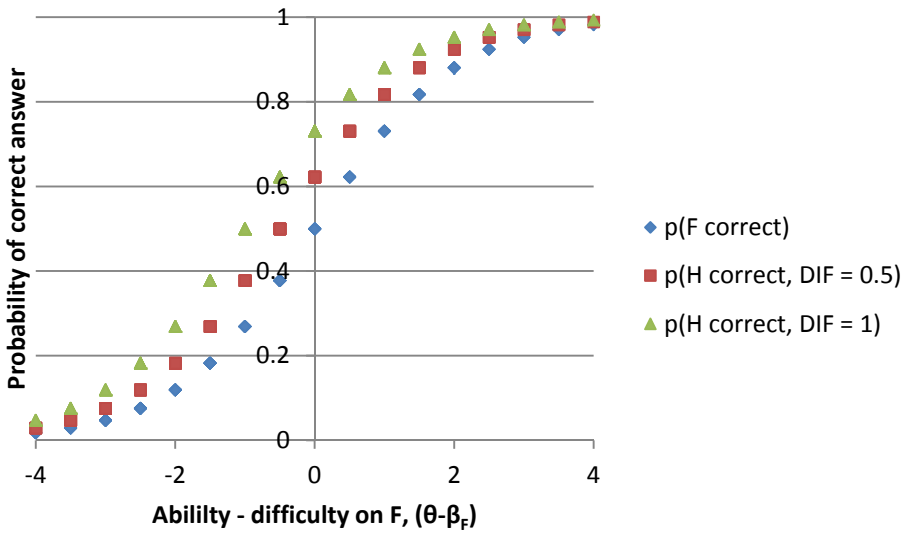

Figure 3 **Probability of correct answer to a dichotomous common item, relative to the difference between ability and difficulty**

Spotting errors in common item identification

Both Figures 1 and 2 could be used to flag items that have been incorrectly identified as common. This is an important practical part of the vertical scaling process as common items are manually identified and keyed, and an error could occur. If DIF was not assessed, such an error might go unnoticed and an inappropriate standard may be applied to link the tiers.

Comparison over 12 papers

The agreement rate between the two methods of detecting DIF across all 12 papers analysed is shown in Table 2. The orthogonal residual is flagging far fewer problem items than the PCM DIF analysis; several of these 'missed' items lie at the extremes of the difficulty plots. This means that the presence of DIF in these items will have little impact on candidates' scores and therefore will not have a large effect on equating results.

[4] Using the Rasch equation: $p(X_{ni} = 1) = \frac{e^{\theta n - \beta i}}{1+e^{\theta n - \beta i}}$, where X_{ni} is the mark (0 or 1) for the n^{th} student on the i^{th} item, θ_n is a parameter describing the ability of the of the n^{th} student, β_i is a parameter describing the difficulty of the i^{th} item, (Rasch, 1960).

At AQA it has been normal practice to assess for DIF using plots like Figure 2; however, it is unusual to find that the equating is much changed by dropping a common item that showed DIF, and, as mentioned above, dropping a common item would only be done after subject experts had been consulted. Although Table 2 shows that there is a relatively small overlap between the two methods of detecting DIF, difficulty plots such as Figure 1 could be a useful, simple tool to use alongside chained equipercentile equating in the future.

Table 2 Detecting DIF: comparison of PCM and difficulty plots over 12 sets of common items

		Orthogonal residual	
		≥0.1	<0.1
PCM	≥0.5	7	23
	<0.5	4	108

Recommendation

In summer 2017, chained equipercentile equating will be used to link reformed tiered papers. To enable the assessment of DIF, the calculation of orthogonal residuals and the creation of difficulty plots (such as Figure 1) should be added to the equating function in the CERP software. It is suggested that any item with an orthogonal residual greater than or equal to 0.1 should be regarded as exhibiting DIF, and consideration should be given as to whether to re-run the equating with that item treated as a non-common item.

References

Harrison, E. A., & Pointer, W. H. (2016). *Mind the gap: problems with vertical scaling in GCSE Mathematics 2015.* Manchester, UK: AQA Centre for Education Research and Practice.

JCQ. (2015). *Report and Recommendations for STAG*. JCQ Awarding Sub-group.

Kolen, M., & Brennan, R. (2004). *Test Equating, Scaling, and Linking*. New York: Springer.

Livingston S. A. (2004). *Equating Test Scores (without IRT).* Educational Testing Service.

Masters, G. N. (1982). A Rasch model for partial credit scoring*. Psychometrika*, *47*(2), 149–174.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen*:* Danish Institute for Educational Research.