
Differential Item Functioning

Ray Adams and Margaret Wu, 29 August 2010

Within the context of Rasch modelling an item is deemed to exhibit differential item functioning (DIF) if the response probabilities for that item cannot be fully explained by the ability of the student and a fixed set of difficulty parameters for that item. Through the use of its multi-faceted modelling capabilities, and more particularly its ability to model interactions between facets, ConQuest provides a powerful set of tools for examining DIF.

In this tutorial; three examples are considered. In the first, ConQuest is used to explore the existence of DIF with respect to gender in a short multiple-choice test. This is a traditional DIF analysis because it is applied to dichotomously scored items and examines DIF between two groups – that is, it uses a binary grouping variable. In the second example DIF is explored when the grouping variable is polytomous—in fact the grouping variable defines eight groups of students. Finally, in the third example DIF in some partial credit items is examined.

EXAMINING GENDER DIFFERENCES IN A MULTIPLE CHOICE TEST

The data used in this first example are the TIMSS data that were described in tutorial five. The files used in this example are:

The files used in this example are:

ex6a.cqc	The command lines used for the first analysis
ex5.dat	The data.
ex6.lab	A file of labels for the items.
ex6a.shw	Selected results from the analysis

The control code for analysing these data is shown in Figure 1 **Error! Reference source not found.**

```
1.   datafile ex5.dat;
2.   title Australian TIMSS Mathematics Data--First
      Six Items--Gender Differences;
3.   formatbook 16 gender 17 level 18 gbyl 19
      responses 20-25;
4.   labels << ex6.lab;
5.   key 134423 ! 1;
6.   model item-gender+item*gender;
7.   estimate !fit=no, stderr=full;
8.   show !table=2 >> ex6a.shw;
9.   plot icc! gins=1:2, overlay=yes, legend=yes;
10.  plot icc! gins=3:4, overlay=yes, legend=yes;
11.  plot icc! gins=5:6, overlay=yes, legend=yes;
12.  plot icc! gins=7:8, overlay=yes, legend=yes;
13.  plot icc! gins=9:10, overlay=yes, legend=yes;
14.  plot icc! gins=11:12, overlay=yes, legend=yes;
```

Figure 1 Control Code For Examining Gender Differences In A Multiple Choice Test

1. The data in `ex5.dat` is to be used.
2. Sets the title.
3. Note that in this format we are reading the explicit variables `book`, `gender`, `level` and the product of `gender` and `level` from columns 16, 17, 18 and 19 respectively.
4. Note that the labels file for this analysis contains labels for `book`, `gender` and `item`.
5. Gives the scoring key.
6. The `model` statement has three terms. These three terms involve two facets, `item` and `gender`. So, as ConQuest passes over the data, it will identify all possible combinations of the `item` and `gender` variables and construct 12 (six items by 2 genders) *generalised* items. The `model` statement requests that ConQuest describes the probability of correct responses to these generalised items using an `item` main effect, a `gender` main effect and an interaction between `item` and `gender`.

The first term will yield a set of item difficulty estimates, the second term will give the mean ability of the male and female students and the third term will give an estimate of the difference in the difficulty of the items for the two gender groups. Note, a negative sign has been used in front of the `gender` term. This ensures that the gender parameters will have the more natural orientation of a higher number corresponding to a higher mean ability.
7. Two options have been included with the `estimate` command. `fit=no`, means that fit statistics will not be computed, and `stderr=full` means that the more time consuming (and more accurate) method will be used to calculate asymptotic standard error estimates for the items. The more accurate method has been chosen for this analysis since the comparison of estimates of some parameters to their standard errors is used in judging whether there is DIF.
8. The `show` command will write table 2 to the file `ex6a.shw`.

9.-14. Plots the item characteristic curves for each of the six items. Because this run of ConQuest uses a multi-faceted model that involves six items and two genders there are actually 12 *generalised* items that are analysed. In the model statement the item facet is mentioned first and the gender facet is mentioned second, as a consequence the gender facet reference cycles fastest in the referencing of generalised items. That is, generalised item one corresponds to item one and gender one; generalised item two corresponds to item one and gender two; generalised item three corresponds to item two and gender one; generalised item four corresponds to item two and gender two; and so on.

Each plot command plots the item characteristic curves for two generalised items. For example the first command plots generalised items one and two, which corresponds to a plot of item one for the two gender groups separately. The `overlay=yes` option results in both item characteristic curves being plotted in the same graph.

RUNNING THE TEST FOR DIF

To run this sample analysis, start the gui version of ConQuest and open the control file

```
Ex6a.cqc
```

Select Run -> Run All. ConQuest will begin executing the statements that are in the file `ex6a.cqc`; and as they are executed they will be echoed in the Output Window. When it reaches the estimation command ConQuest will begin fitting a multi-faceted model to the dichotomous data. This analysis will converge in 33 iterations and the item parameter estimates will be written to the file `ex6a.shw`.

Differential Item Functioning

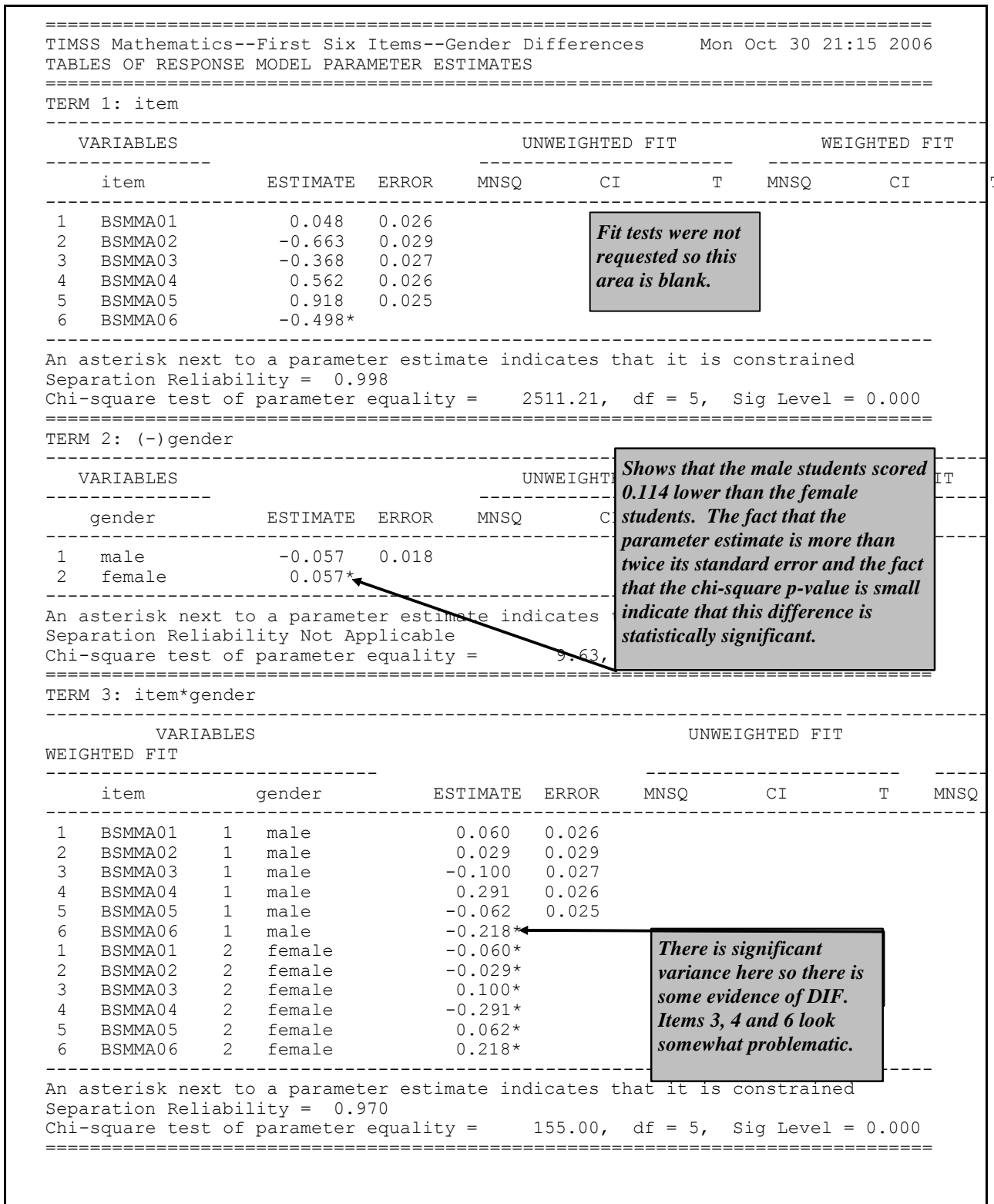


Figure 2 Parameter Estimates For DIF Examination In A Multiple Choice Test

The contents of ex6a.shw are shown in Figure 2. The figure contains three tables, one for each of the terms in the model statement. The first table shows the item difficulty parameter estimates for each of the six items.

The second table shows the estimates for the gender differences in ability estimates. A negative sign was used for the gender term in the item response model so these results indicate that the male students have performed more poorly than the female students. The actual parameter estimate for the male students is three times larger than its standard error estimate so the difference between the male and female means is obviously significant. The chi-square value of 9.63 on one degree of freedom is consistent with this finding. The conclusion that can be drawn here is that the male mean performance is lower than that of the females, this DOES NOT indicate differential item functioning. Further, the estimated difference of 0.114 is small at just over 10% of a student standard deviation.¹

Note: *By including the main effect, gender, in the item response model, estimates of the mean scores for male and female students has been obtained. An alternative approach that would have achieved an identical result would have been to place the gender variable in the population model. It would not be appropriate to include gender in both the item response and the population models since this would make the model unidentified.*

The third table gives the interaction between the item and gender facets. The estimate of 0.060 for item BSMMA01 and males indicates that 0.060 must be added to the difficulty of this item for male students, similarly -0.060 must be added for the females. That is, female students found this item to be relatively easier than did the males. The results in this table show that three items (BSMMA03, BSMMA05 and BSMMA06) are relatively easier for males than females, two items (BSMMA01 and BSMMA04) are relatively easier for females than males, and one item (BSMMA02) has the same difficulty. The significant chi-square (155.00, df=5) also shows the existence of DIF.

Warning: *The current version of ConQuest assumes independence between the parameter estimates when computing the chi-square test of parameter equality.*

While this analysis has shown the existence of DIF in these items it is the magnitude of that DIF that will determine if the effect of that DIF is of substantive importance. For example, the first item (BSMMA01) is significantly more difficult for males than females but the difference estimate is just 0.12 logits. If all of the items exhibited DIF of this magnitude it would shift the male ability distribution by just over 10% of a student standard deviation. With just one item having this DIF, the effect is much smaller. The fourth item (BSMMA04) exhibits much more DIF. In fact if all of the items in the test had behaved like this item the estimated mean score for the males would be 0.582 logits lower than that of the females, that is more than 50% of a student standard deviation.

¹ The standard deviation is around 1.1. The results reported here should not be extrapolated to the Australian TIMSS data. The significance testing done here does not take account of the design effects that exist in TIMSS due to the cluster sampling that was used, further they are based on a random selection of half of the TIMSS data set.

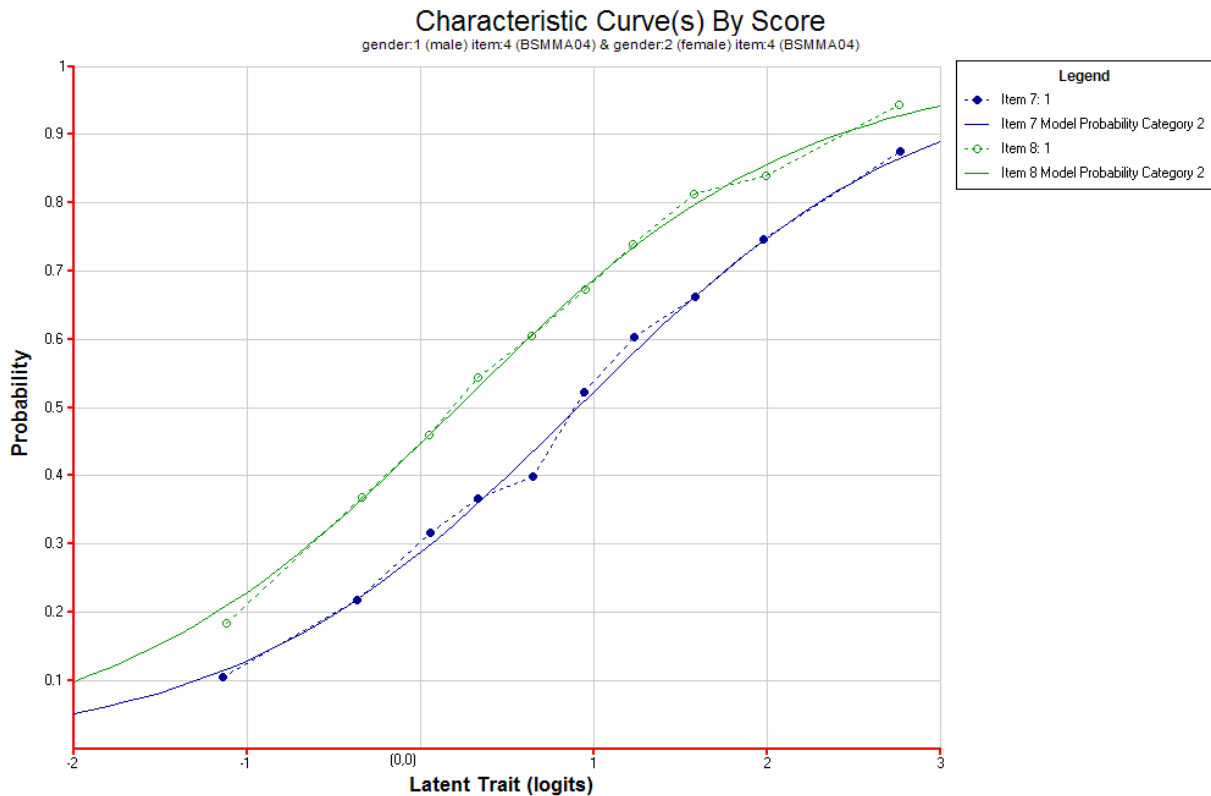


Figure 3 Item Characteristic Curves for Generalised Items Seven and Eight (Item 4, Males and Females)

Figure 3 shows the item characteristic curves for Item 4 for males and females separately. The blue curves (darker ones) are for males, and the green curves (lighter ones) are for females. It can be seen that, given a particular ability level, the probability of being successful on this item is higher for females than for males, That is, females find this item easier than males.

EXAMINING DIF WHEN THE GROUPING VARIABLE IS POLYTOMOUS

ConQuest can also be used to examine DIF when the grouping variable is polytomous, rather than dichotomous, as is the case with gender. In the TIMSS design the test items were allocated to eight different testing booklets and students were allocated one of the eight booklets at random. One way of testing whether the rotation scheme was implemented successfully is to estimate the mean ability estimates for the students who were assigned each booklet and to see if there is any evidence of DIF across the booklets.

The files that we will use in this example are:

- | | |
|----------|--|
| ex6b.cqc | The command lines used for the second analysis |
| ex5.dat | The data. |
| ex6.lab | A file of labels for the items. |
| ex6b.shw | Selected results from the analysis |

Figure 4 Control Code For Examining Booklet Effect In A Multiple Choice Test

```
1. datafile ex5.dat;
2. title Australian TIMSS Mathematics Data--First
   Six Items--Booklet Differences;
3. format book 16 gender 17 level 18 gbyl 19
   responses 20-25;
4. labels << ex6.lab;
5. key 134423 ! 1;
6. model item+book+item*book;
7. estimate !stderr=full;
8. show !table=2 >> ex6b.shw;
```

The contents of the control file, `ex6b.cqc`, is shown in Figure 4. The only command that is different here to that used in Figure 1 is the model, in which the variable `book` rather than `gender` is used. After running this analysis using the same procedures as described for previous examples, the file `ex6b.shw` will be produced, the contents of which are shown in Figure 5.

This figure shows that there is no statistically significant *book effect* and that there is no between booklet DIF.

Differential Item Functioning

Australian TIMSS Mathematics Data--First Six Items--BookletTue Oct 31 12:30 2006
 TABLES OF RESPONSE MODEL PARAMETER ESTIMATES

=====

TERM 1: item

VARIABLES		UNWEIGHTED FIT				WEIGHTED FIT		
item	ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1 BSMMA01	0.051	0.026	0.89	(0.97, 1.03)	-6.7	0.92	(0.98, 1.02)	-6.4
2 BSMMA02	-0.662	0.028	1.02	(0.97, 1.03)	1.1	1.01	(0.97, 1.03)	0.5
3 BSMMA03	-0.375	0.027	0.94	(0.97, 1.03)	-3.6	0.96	(0.97, 1.03)	-2.8
4 BSMMA04	0.584	0.025	0.99	(0.97, 1.03)	-0.3	0.99	(0.98, 1.02)	-0.6
5 BSMMA05	0.912	0.025	1.14	(0.97, 1.03)	7.8	1.10	(0.98, 1.02)	8.7
6 BSMMA06	-0.509*		1.07	(0.97, 1.03)	4.1	1.05	(0.97, 1.03)	2.9

An asterisk next to a parameter estimate indicates that it is constrained
 Separation Reliability = 0.99
 Chi-square test of parameter equality = 3.48, df = 5, Sig Level = 0.000

=====

TERM 2: book

VARIABLES		UNWEIGHTED FIT				WEIGHTED FIT		
book	ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1 book1	0.035	0.047	1.01	(0.91, 1.09)	0.5	0.91	(0.93, 1.07)	-2.8
2 book2	-0.000	0.048	1.01	(0.91, 1.09)	1.0	1.00	(0.91, 1.09)	-0.0
3 book3	-0.026	0.048	1.01	(0.91, 1.09)	-1.5	0.96	(0.92, 1.08)	-1.0
4 book4	-0.058	0.048	0.97	(0.91, 1.09)	-0.9	0.96	(0.94, 1.06)	-1.2
5 book5	-0.030	0.048	0.98	(0.91, 1.09)	1.0	1.03	(0.94, 1.06)	1.0
6 book6	0.038	0.048	0.96	(0.90, 1.10)	1.9	1.05	(0.92, 1.08)	1.2
7 book7	0.019	0.049	0.99	(0.90, 1.10)	-2.3	0.92	(0.93, 1.07)	-2.1
8 book8	0.022*		0.98	(0.90, 1.10)	-0.1	1.02	(0.91, 1.09)	0.4

An asterisk next to a parameter estimate indicates that it is constrained
 Separation Reliability = 0.000
 Chi-square test of parameter equality = 3.48, df = 7, Sig Level = 0.837

=====

TERM 3: item*book

VARIABLES			UNWEIGHTED FIT				WEIGHTED FIT		
item	book	ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1 BSMMA01	1 book1	0.094	0.068	0.87	(0.91, 1.09)	-2.8	0.91	(0.93, 1.07)	-2.8
2 BSMMA02	1 book1	-0.026	0.074	1.05	(0.91, 1.09)	1.0	1.00	(0.91, 1.09)	-0.0
3 BSMMA03	1 book1	-0.026	0.071	0.93	(0.91, 1.09)	-1.5	0.96	(0.92, 1.08)	-1.0
4 BSMMA04	1 book1	0.008	0.066	0.96	(0.91, 1.09)	-0.9	0.96	(0.94, 1.06)	-1.2
5 BSMMA05	1 book1	-0.058	0.066	1.05	(0.91, 1.09)	1.0	1.03	(0.94, 1.06)	1.0
6 BSMMA06	1 book1	0.008*		1.09	(0.91, 1.09)	1.9	1.05	(0.92, 1.08)	1.2
1 BSMMA01	2 book2	-0.047	0.069	0.89	(0.90, 1.10)	-2.3	0.92	(0.93, 1.07)	-2.1
2 BSMMA02	2 book2	0.008	0.075	0.99	(0.90, 1.10)	-0.1	1.02	(0.91, 1.09)	0.4
3 BSMMA03	2 book2	0.032	0.072	0.94	(0.90, 1.10)	-1.3	0.95	(0.92, 1.08)	-1.1
4 BSMMA04	2 book2	-0.035	0.067	0.92	(0.90, 1.10)	-1.7	0.94	(0.94, 1.06)	-2.0
5 BSMMA05	2 book2	0.077	0.068	1.09	(0.90, 1.10)	1.9	1.06	(0.94, 1.06)	1.8
6 BSMMA06	2 book2	-0.035*		1.09	(0.90, 1.10)	1.7	1.05	(0.91, 1.09)	1.2
1 BSMMA01	8 book8	0.089*		0.92	(0.90, 1.10)	-1.7	0.93	(0.93, 1.07)	-2.1
2 BSMMA02	8 book8	0.022*		1.07	(0.90, 1.10)	1.5	1.04	(0.91, 1.09)	0.8
3 BSMMA03	8 book8	-0.087*		0.97	(0.90, 1.10)	-0.6	0.99	(0.92, 1.08)	-0.2
4 BSMMA04	8 book8	-0.022*		1.02	(0.90, 1.10)	0.4	1.01	(0.94, 1.06)	0.4
5 BSMMA05	8 book8	-0.018*		1.16	(0.90, 1.10)	3.1	1.11	(0.94, 1.06)	3.5
6 BSMMA06	8 book8	0.017*		1.07	(0.90, 1.10)	1.4	1.04	(0.92, 1.08)	0.9

An asterisk next to a parameter estimate indicates that it is constrained
 Separation Reliability = 0.240
 Chi-square test of parameter equality = 44.95, df = 35, Sig Level = 0.121

=====

Figure 5 Parameter Estimates For DIF Examination Across Booklets

DIF FOR POLYTOMOUS ITEMS

As a final example on DIF, a set of polytomous items is examined. The data were collected by Adams, Doig and Rosier (1991) as a part of their study of science achievement. The set of items

that are analysed formed an instrument that assessed students' understanding of Force and Motion.

The files used in this example are:

ex6c.cqc	The command lines used for the third set of analyses
ex6.dat	The data.
ex6c.lab	The variable labels for the items on the test.
ex6c.shw	The results of an analysis that includes gender by step interactions.
ex6d.shw	The results from an analysis that does not include gender by step interactions.

The control code for this example is shown in Figure 6. Note that in this case the control code will actually run two ConQuest analyses. The control code listed in this figure is very similar to that described in the earlier examples in this tutorial. Only the distinguishing aspects of the code in Figure 6 are commented upon.

```

1.  datafile ex6.dat;
2.  format responses 10-18 grade 118 gender 119 !
      tasks (9);
3.  set warnings=no;
4.  model tasks - gender + tasks*gender +
      gender*tasks*step;
5.  labels << ex6c.lab;
6.  estimate !fit=no,stderr=full;
7.  show !table=1:2 >>ex6c.shw;
8.  plot expected!gins=1:2,overlay=yes,legend=yes;
9.  reset;
10. datafile ex6.dat;
11. format responses 10-18 grade 118 gender 119 !
      tasks (9);
12. set warnings=no;
13. model tasks-gender+tasks*gender+tasks*step;
14. labels << ex6c.lab;
16. estimate !fit=no,stderr=full;
17. show !table=1:2 >>ex6d.shw;
18. plot expected!gins=1:2,overlay=yes,legend=yes;

```

Figure 6 Control Code for Examining DIF In Partial Credit Items

4. This model includes four terms. Two main effects, tasks and gender, give the difficulty of each of the tasks and the means of the two gender groups. The interaction $tasks*gender$ models the variation in difficulty of the task between the two genders and finally the $gender*tasks*step$ term models differing step structures for each task and gender.

Extension: In this example randomly chosen students from both an upper and lower grade responded to all of the tasks so the use of grade as a regressor is not necessary to produce consistent estimates of the item response model parameters.

If the sub-samples of students who respond to specific test tasks were systematically different in their latent ability distribution then the use of a regressor will be necessary to produce consistent parameter estimates for the item response model (see Mislevy and Sheehan, 1989).

9. The `reset` command separates sets of analyses to be run.

12. This `model` command is similar to the previous one in that it has four terms. The difference is that the final term does not include variation between males and females in the task's step structure. Comparing the fit of this model to the model given by line 4, we can assess the need for a step structure that is different for male and female students.

After this analysis is run using the same procedures as described for previous examples, the files `ex6c.shw` and `ex6d.shw` will be produced. An extract of `ex6c.shw` is given in Figure 7, it shows that there is no difference between the overall performance of male and female students and that there is no interaction between gender and task difficulty. In this figure the parameter estimates for the term `gender*tasks*step` are not shown because the easiest way to test whether the step structure is the same for the male and female students is to compare the deviance of the two models that were fitted by the code in Figure 6 **Error! Reference source not found.**

The results reported in Figure 7 show that the model with a step structure that is invariant to gender does not fit as well as the model with a step structure that varies with gender. The conclusion that can be drawn from these analyses is that while the overall male and female performance is equivalent, as are the difficulty parameters for each of the tasks it appears that male and female students have differing step structures. A closer examination of the difference in the step structures between male and female students would appear to be required.

To illustrate the differences between these two models, the expected score curves have been plotted for the first two generalised items for each model. The plots are shown in Figure 8. The first plot shows the expected score curves when a different step structure is used for male and female students, while the second plot shows the expected score curves when a common step structure is used. In the second plots the curves are parallel, in the sense that they have the same shape but are just displaced on the horizontal axes. In the first plots the expected curves take a different shape, and in fact cross.

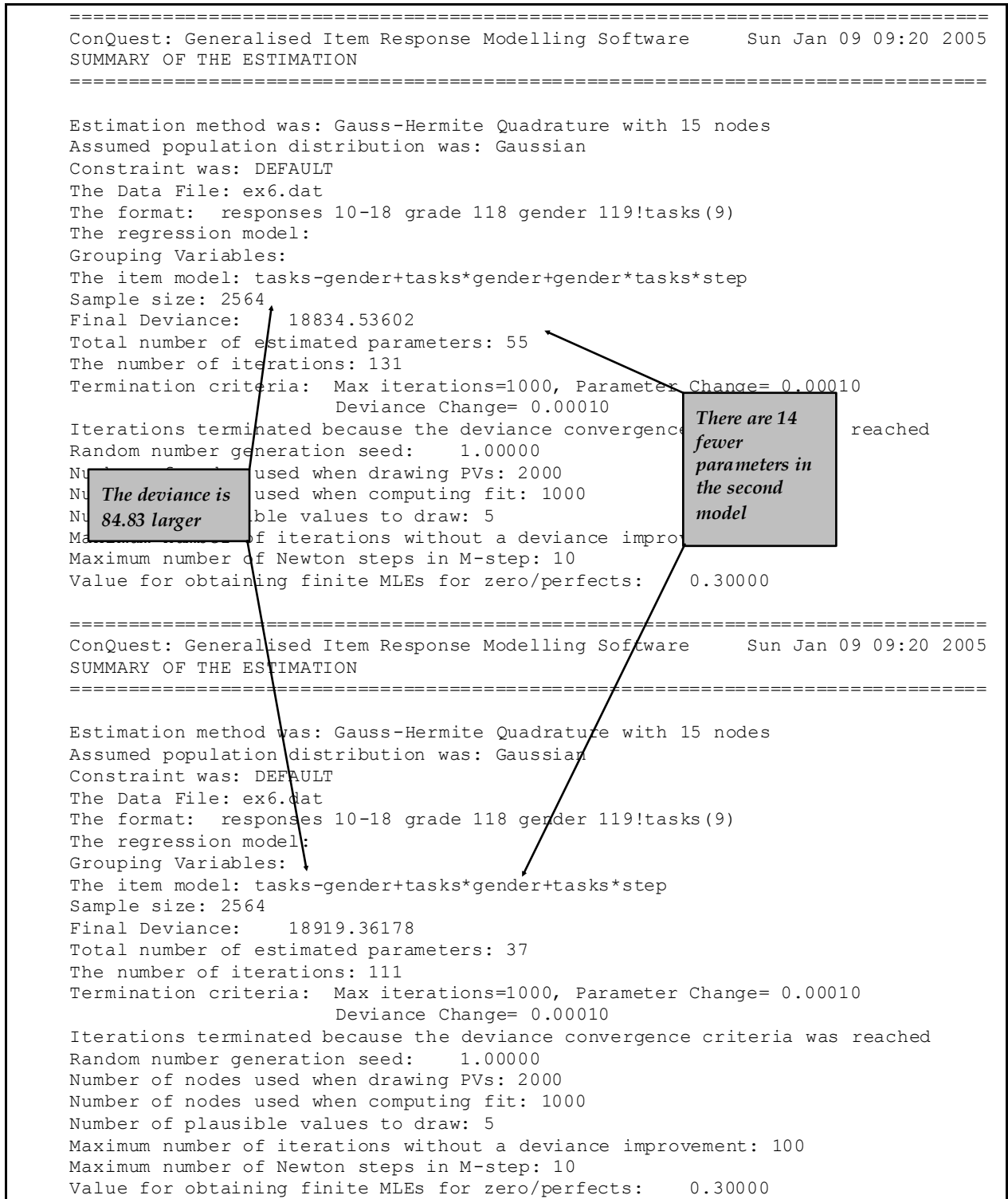


Figure 7 The Summary Tables For The Two Polytomous DIF Analyses

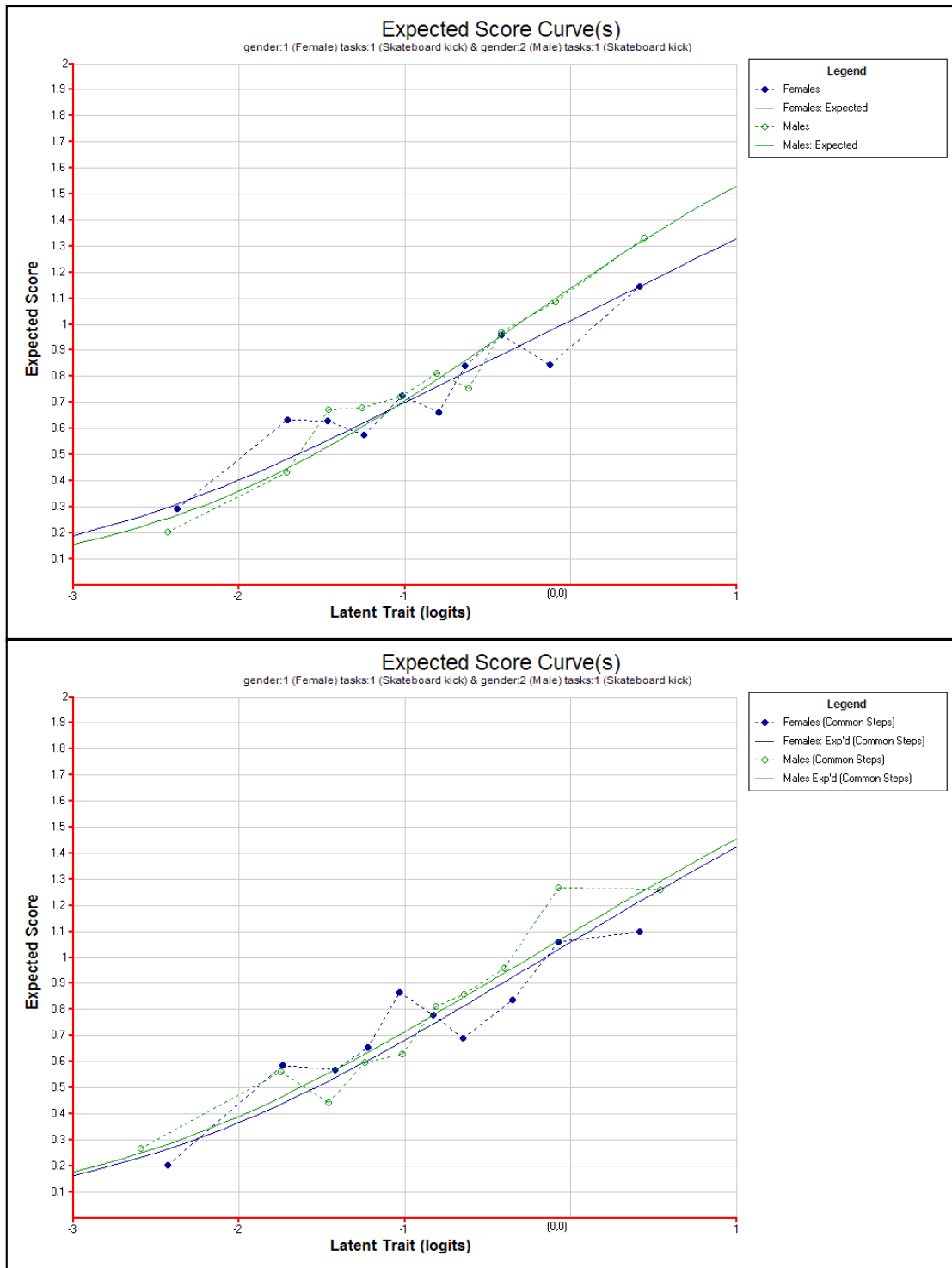


Figure 8 Output From Analysis of DIF In Polytomous Items

SUMMARY

In this tutorial we have illustrated how ConQuest can be used to examine DIF with dichotomous items and polytomous items, and how DIF can be explored where the grouping variable is polytomous.

Some key points covered in this tutorial are:

- Modelling DIF can be done through adding an item-by-facet interaction term in the Model statement.
- Item characteristic curves can be plotted with the overlay option.
- A comparison of model fit can be carried out using the deviance statistic.
- Expected score curves are useful for polytomous items.
- Different steps structures can be specified using the model statement.

REFERENCES

Adams, R. J., Doig, B. A., and Rosier, M. 1991. *Science Learning in Victorian Schools*. Hawthorn, Victoria: Australian Council for Educational Research.