The Analysis of Rater Effects

Ray Adams and Margaret Wu. 22 August 2010

Item response models, such as simple logistic, rating scale and partial credit assume that the observed responses result from the two-way interaction between the agents of measurement[1] and the objects of measurement[2]. With the increasing importance of performance assessment, Linacre ([1989] 1994) recognised that the responses that are gathered in many contexts do not result from the interaction between an object and a single agent: the agent is often a composite of more fundamental subcomponents.[3] Consider, for example, the assessment of writing where a stimulus is presented to a student, the student prepares a piece of writing, and then a rater makes a judgment about the quality of the writing performance. Here, the object of measurement is clearly the student; but the agent is a combination of the rater who makes the judgment and the stimulus that serves as a prompt for the student's writing. The response that is analysed by the item response model is influenced by the characteristics of the student, the characteristics of the stimulus, and the characteristics of the rater. Linacre ([1989] 1994) would label this a three-faceted measurement context, the three facets being the student, the stimulus and the rater.

Using an extension of the partial credit model to this multifaceted context, Linacre ([1989] 1994) and others have shown that item response models can be used to identify raters who are harsher or more lenient than others, who exhibit different patterns in the way they use rating schemes, and who make judgments that are inconsistent with judgments made by other raters. This tutorial describes how ConQuest can fit a multifaceted measurement model to analyse the characteristics of a set of 16 raters who have rated a set of writing tasks using two criteria.

FITTING A MULTIFACETED MODEL

The data that we are analysing are the ratings of 8296 Year 6 students' responses to a single writing task. The data were gathered as part of a study reported in Congdon and McQueen (1997). Each of the 8296 students' writing scripts was graded by two raters, randomly chosen from a set of 16 raters; and the second rating for each script was performed blind. The random allocation of scripts to the raters, in conjunction with the very large number of scripts, resulted in links between all raters being obtained. When assessing the scripts, each rater was required to provide two ratings, one labelled OP (overall performance) and the other TF (textual features).[4] The rating of both the OP and TF was undertaken against a six-point scale, with the labels G, H, I, J, K and L used to indicate successively superior levels of performance. For a

---

1 The agents of measurement are the tools that are used to stimulate responses. They are typically test items or, more generally, assessment tasks.

2 The object of measurement is the entity that is to be measured, most commonly a student, a candidate or a research subject.

3 Fischer (1973) recognised that items could be described by more fundamental parameters when he proposed the linear logistic test model. Linacre ([1989] 1994) extended the model to the polytomous case and recognised that the more fundamental components could be raters and such.

4 OP (overall performance) is a judgment of the task fulfilment, particularly in terms of appropriateness for purpose and audience, conceptual complexity, and organisation of the piece. TF (textual features) focuses on control and effective use of syntactic features, such as cohesion, subordination, and verb forms, and other linguistic features, such as spelling and punctuation.

small number of scripts, ratings of this nature could not be made; and the code N was used to indicate this occurrence.

The files used in this sample analysis are:

`ex3a.cqc`	The command statements.
`ex3.dat`	The data.
`ex3a.shw`	The results of the multifaceted analysis.
`ex3a.itn`	The results of the traditional item analyses.

(The last two files are created when the command file is executed.)

The data were entered into the file `ex3.dat`, using one line per student. Rater identifiers (of two characters in width) for the first and second raters who rated the writing of each student are entered in columns 17 and 18 and columns 19 and 20, respectively. Each of the two raters produced an OP and a TF rating for the script. The OP and TF ratings made by the first rater have been entered in columns 21 and 22, and the OP and TF ratings made by the second rater have been entered in columns 25 and 26. The command file for fitting one possible multifaceted model to these data is shown in Figure 1.

```
1.  Title Rater Effects Model One;
2.  datafile ex3.dat;
3.  format rater 17-18 rater 19-20 responses 21-22
       responses 25-26 ! criteria(2);
4.  codes G,H,I,J,K,L;
5.  score (G,H,I,J,K,L) (0,1,2,3,4,5);
6.  labels 1 OP !criteria;
7.  labels 2 TF !criteria;
8.  model rater + criteria + step;
9.  estimate;
10. show ! estimates=latent >> ex3a.shw;
11. itanal >> ex3a.itn;
```

Figure 1 Sample Command File for Multifaceted Data

1. Gives a title for the analysis. The text supplied after the `title` command will appear on the top of any printed ConQuest output.

2. Indicates the name and location of the data file.

3. Multifaceted data can be entered into data sets in many ways. Here, two sets of ratings for each student have been included on each line in the data file, and explicit rater codes have been used to identify the raters. For each of the raters, there is a matching pair of ratings (one for OP and one for TF). The OP and TF ratings are implicitly identified by the columns in which the data are entered. The ConQuest `format` statement is very flexible and can cater for many alternative data specifications. In this `format` statement, you will notice that `rater` is used twice. The first use indicates the column location of the rater code for the first rater, and the second use indicates the column location of the rater code for the second rater. This is followed by two variables indicating the location of the responses (referred to as response blocks). Each response block is two characters wide; and since the default width of a response is one column, each response block refers to two responses, an OP and a TF rating. The first response block (columns 21 and 22) will be associated with the first rater, and the second response block (columns 25 and 26) will be associated with the second rater.

This `format` statement also includes an option, `criteria(2)`, which assigns the variable name `criteria` to the two responses that are implicitly identified by each response block. If this option had been omitted, the default variable name for the responses would be `item`.

This `format` statement spans two lines in the command file. Command statements can be 1023 characters in length and can cover any number of lines in a command file. The semi-colon (;) is the separator between statements, not the `return` or `new line` characters.

4. The `codes` statement restricts the list of valid response codes to G, H, I, J, K, and L. All other responses will be treated as missing-response data.

5. The `score` statement assigns score levels to each of the response categories. Here, the left side of the `score` argument shows the six valid codes defined by the `codes` statement, and the right side gives six matching scores. The six distinct codes on the left indicate that the item response model will model six categories for each item; the scores on the right are the scores that will be assigned to each category.

NOTE:	***ConQuest makes an important distinction between response categories and response levels (or scores). The number of item response categories that will be modelled by ConQuest is determined by the number of unique codes that exist after all recodes have been performed. ConQuest requires a score for each response category. This can be provided via the `score` statement. Alternatively, if the `score` statement is omitted, ConQuest will treat the recoded responses as numerical values and use them as scores. If the recoded responses are not numerical values, an error will be reported.***

6.-7. In the previous sample analyses, variable labels were read from a file. Here the `criteria` facet contains only two levels (the OP and TF ratings), so the labels are given in the command file using `labels` command syntax. These `labels` statements have two arguments. The first argument indicates the level of the facet to which the label is to be assigned, and the second argument is the label for that level. The option gives the facet to which the label is being applied.

8. The `model` statement here contains three terms; `rater`, `criteria` and `step`. This `model` statement indicates that the responses are to be modelled with three sets of parameters: a set of rater severity parameters, a set of criteria difficulty parameters, and a set of parameters to describe the step structure of the responses.

9. The `estimate` statement initiates the estimation of the item response model.

10. The `show` statement produces a display of the item response model parameter estimates and saves them to the file `ex3a.shw`. The option `estimates=latent` requests that the displays include an illustration of the latent ability distribution.

11. The `itanal` statement produces a display of the results of a traditional item analysis. As with the `show` statement, we have redirected the results to a file (in this case, `ex3a.itn`).

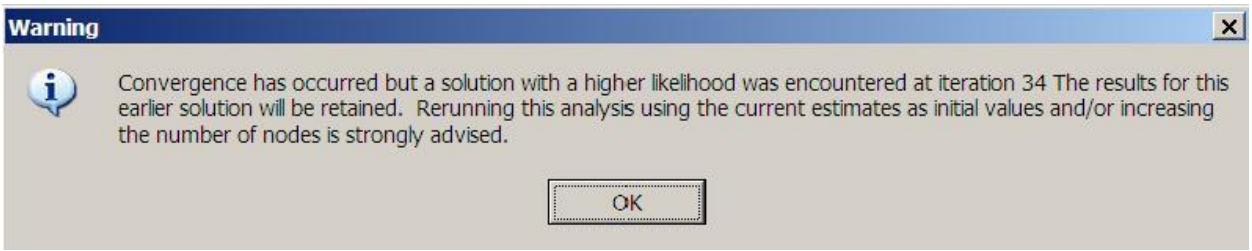
EXTENSION:	*The* **`model`** *statement in this sample analysis includes main effects only. An interaction term* **`rater*criteria`** *could be added to model variation in the difficulty of the criteria across the raters. Similarly, the model specifies a single step-structure for all rater and criteria combinations. Step structures that were common across the criteria but varied with raters could be modelled by using the term* **`rater*step`***, step structures that were common across the raters but varied with criteria could be modelled by using the term* **`criteria*step`***, and step structures that varied with rater and criteria combinations could be modelled by using the term* **`rater*criteria*step`***.*

RUNNING THE MULTIFACETED SAMPLE ANALYSIS

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

```
ex3a.cqc
```

Select Run -> Run All. ConQuest will begin executing the statements that are in the file `ex3a.cqc`; and as they are executed, they will be echoed on the screen. When ConQuest reaches the `estimate` statement, it will begin fitting the multifaceted model to the data; and as it does, it will report on the progress of this estimation. Due to the large size of this data file, ConQuest will take some time to perform this analysis. After 65 iterations, ConQuest reports a warning message:



As the scores of the writing test spread students far apart, as indicated by the estimated variance of the ability distribution (5.7 logits), this suggests that more nodes to cover the ability range are required in the estimation process.

To re-run ConQuest with more nodes during the estimation, modify the `estimate` command as follows:

```
9.  Estimate ! nodes=30;
```

The default number of nodes is 15. The above `estimate` command requests ConQuests to use 30 nodes to cover the ability range.

Re-run ConQuest by selecting Run -> Run All from the menu. This time, ConQuest no longer reports a warning for convergence problems.

After the estimation is complete, the two statements that produce output (`show` and `itanal`) will be processed. The results of the `show` statement can be found in the file `ex3a.shw`, and the results of the `itanal` statement can be found in the file `ex3a.itn`. On this occasion, the `show` statement will produce six tables.

From Figure 2, we note that there were 16 raters and that the severity ranges from a high of 0.977 logits for rater 14 (the first rater in the table) to a low of -1.292 for rater 19 (the fourth rater in the table). This is a range of 2.123, which appears quite large when compared to the standard deviation of the latent distribution, which is estimated to be 2.37 (the square root of the variance that is reported in the third table (the population model) in `ex3a.shw`). That means that ignoring the influence of the severity of the raters may move a student's ability estimate by as much as one standard deviation of the latent distribution. We also note that, with this model, the raters do not fit particularly well. The high mean squares (and corresponding positive t values) suggest quite a bit of unmodelled noise in the ratings.

```
Rater Effects Model One                                    Tue Oct 03 16:42 2006
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
TERM 1: rater
   VARIABLES                                  UNWEIGHTED FIT              WEIGHTED FIT
    rater      ESTIMATE   ERROR^    MNSQ        CI          T     MNSQ       CI          T
 1   14          0.977    0.029     1.15 ( 0.92, 1.08)   3.3    1.18 ( 0.91, 1.09)   4.0
 2   17          0.125    0.029     1.32 ( 0.91, 1.09)   6.3    1.34 ( 0.91, 1.09)   6.7
 3   18         -0.088    0.031     1.82 ( 0.90, 1.10)  13.2    1.86 ( 0.90, 1.10)  13.5
 4   19         -1.292    0.028     1.30 ( 0.91, 1.09)   6.3    1.32 ( 0.91, 1.09)   6.6
 5   24          0.639    0.029     1.56 ( 0.91, 1.09)  10.8    1.58 ( 0.91, 1.09)  11.0
 6   38         -0.113    0.030     1.11 ( 0.91, 1.09)   2.3    1.13 ( 0.90, 1.10)   2.5
 7   67          0.538    0.029     1.17 ( 0.92, 1.08)   3.7    1.19 ( 0.91, 1.09)   4.2
 8   70          0.111    0.029     1.13 ( 0.91, 1.09)   2.8    1.13 ( 0.91, 1.09)   2.8
 9   73         -0.003    0.028     1.14 ( 0.92, 1.08)   3.2    1.13 ( 0.92, 1.08)   3.0
10   74         -0.221    0.027     1.34 ( 0.92, 1.08)   8.1    1.33 ( 0.92, 1.08)   7.7
11   78                                    1.09)   4.9    1.27 ( 0.91, 1.09)   5.6
12   79                                    1.08)   1.8    1.09 ( 0.92, 1.08)   2.0
13   8                                     1.09)   2.5    1.13 ( 0.91, 1.09)   2.8
14   85                                    1.08)   8.9    1.45 ( 0.92, 1.08)   9.5
15   89                                    1.10)   3.6    1.22 ( 0.90, 1.10)   3.9
16   93                                    1.09)   4.3    1.23 ( 0.91, 1.09)   4.8
An asterisk next to a parameter estimate indicates that it is
Separation Reliability =   0.997
Chi-square test of parameter equality =   4921.30,   df = 15,
^ Quick standard errors have been used
```

*This part of the table is for the **rater** term.*

There are 16 raters. Labels for the raters were not provided, so the values listed here are the values that were found in the data file.

The fit statistics for most of the raters are larger than one by a substantial amount.

Figure 2 Parameter Estimates for Rater Severity

In Figure 3, we note that the OP and TF difficulty estimates are very similar, differing by just 0.178 logits. This difference is significant but very small. The mean square fit statistics are less than one, suggesting that the criteria could have unmodelled dependency.

```
TERM 2: criteria
   VARIABLES                                  UNWEIGHTED FIT              WEIGHTED FIT
   criteria    ESTIMATE   ERROR^    MNSQ        CI          T     MNSQ       CI          T
 1   OP          0.089    0.010     0.97 ( 0.97, 1.03)  -2.1    0.97 ( 0.97, 1.03)  -1.8
 2   TF         -0.089*   0.010     1.01 ( 0.97, 1.03)   0.8    0.99 ( 0.97, 1.03)  -0.4
An asterisk next to a parameter estimate indicates that it is c
```

*This part of the table is for the **criteria** term.*

The criteria labels are OP and TF.

*There are only two criteria, so that effectively means one **criteria** difficulty estimate, since the average must be zero.*

The fit is less than one, suggesting dependency between these criteria.

Figure 3 Parameter Estimates for the Criteria

Figure 4 shows the step parameter estimates. The fit here is not very good, particularly for steps 1 and 4, suggesting that we should model step structures that interact with the facets. It is pleasing to note that the estimates for the steps themselves are ordered and well separated.

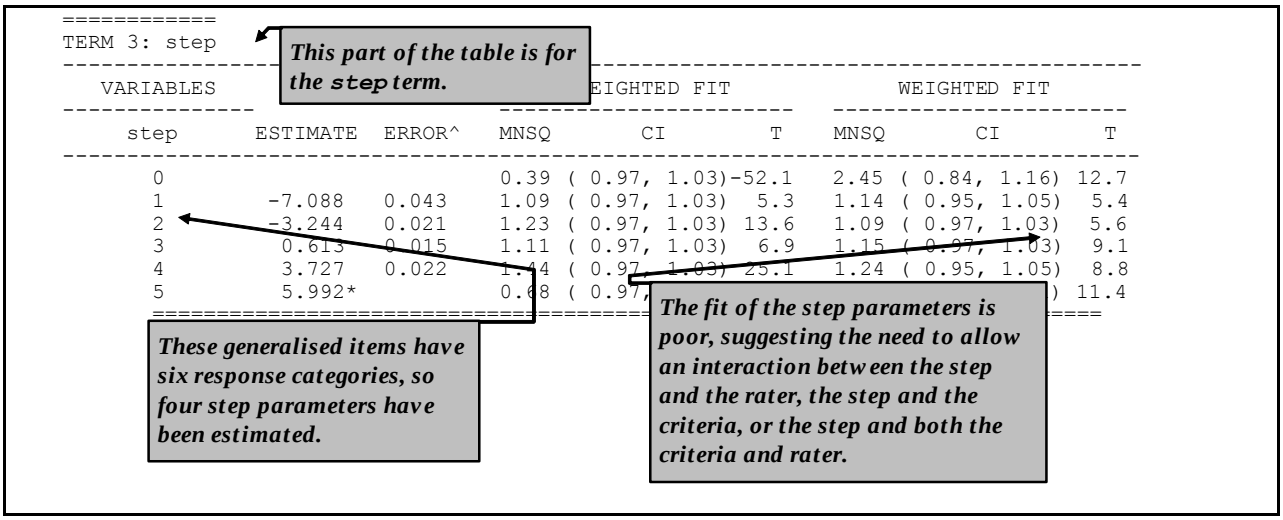


Figure 4 Parameter Estimates for the Steps

Figure 5 is the map of the parameter estimates that is provided in `ex3a.shw`. The map shows how the variation between raters in their severity is large relative to the difference in the difficulty of the two tasks. It also shows that the rater severity estimates are well centred for the estimated ability distribution.

The file `ex3a.itn` contains basic traditional statistics for this multifaceted analysis, extracts of which are shown in Figures 6 and 7.

In this analysis, the combination of the 16 raters and two criteria leads to 32 *generalised items*.[5] The statistics for each of these generalised items is reported in the file `ex3a.itn`. Figure 6 shows the statistics for the last generalised item, which is the combination of rater 93 (the sixteenth rater) and criterion TF (the second criterion). For this generalised item, the total number of students rated by this rater on this criteria is shown (in this case, 1002); and an index of discrimination (the correlation between students' scores on this item and their total score) is shown (in this case, 0.87). This discrimination index is very high, but it should be interpreted with care since only four generalised items are used to construct scores for each student. Thus, a student's score on this generalised item contributes 25% to their total score.

For each response category of this generalised item, the number of observed responses is reported, both as a count and as a percentage of the total number of responses to this generalised item. The point-biserial correlations that are reported for each category are computed by constructing a set of dichotomous indicator variables, one for each category. If a student's response is allocated to a category for an item, then the indicator variable for that category will be coded to 1; if the student's response is not in that category, it will be coded to 0. The point biserial is then the correlation between the indicator variable and the student's total score. It is desirable for the point biserials to be ordered in a fashion that is consistent with the category scores. However, sometimes point biserials are not ordered when a very small or a very large proportion of the item responses are in one category. This can be seen in Figure 6, where only seven of the 1002 cases have responses in category G.

[5] Generalised item is the term that ConQuest uses to refer to each of the unique combinations of the facets that are the agents of measurements.

```
============================================================================
Rater Effects Model One                                  Fri Oct 18 11:25:34
MAP OF LATENT DISTRIBUTIONS AND RESPONSE MODEL PARAMETER ESTIMATES
============================================================================
                 Terms in the Model (excl Step terms)
                        +rater               +criteria
----------------------------------------------------------------------
                X|                          |                       |
                X|                          |                       |
               XX|                          |                       |
   4          XXX|                          |                       |
              XXX|                          |                       |
              XXX|                          |                       |
   3         XXXX|                          |                       |
           XXXXXX|                          |                       |
           XXXXXX|                          |                       |
   2       XXXXXX|                          |                       |
         XXXXXXXX|                          |                       |
        XXXXXXXXX|                          |                       |
   1    XXXXXXXXX|1                         |                       |
       XXXXXXXXXX|5                         |                       |
        XXXXXXXXX|7 15                      |                       |
   0      XXXXXXX|2 8 9 13 14               |1                      |
         XXXXXXXX|3 6 10 12                 |2                      |
         XXXXXXXX|11 16                     |                       |
  -1     XXXXXXXX|                          |                       |
       XXXXXXXXXX|4                         |                       |
          XXXXXXX|                          |                       |
           XXXXXX|                          |                       |
  -2         XXXX|                          |                       |
             XXXX|                          |                       |
             XXXX|                          |                       |
  -3           XX|                          |                       |
               XX|                          |                       |
               XX|                          |                       |
  -4           XX|                          |                       |
                X|                          |                       |
                X|                          |                       |
  -5            X|                          |                       |
======================================================================
Each 'X' represents  51.4 cases
```

Figure 5 Map of the Parameter Estimates for the Multifaceted Model

```
Item 32
------
Rater:16 (93) criteria:2 (TF)
Cases for this item   1002   Discrimination  0.87
Item Threshold(s):    -7.61 -3.75  0.09  3.18  5.58
Item Delta(s):     -7.59 -3.75  0.11  3.23  5.49
-----------------------------------------------------------------------------
Label    Score     Count   % of tot  Pt Bis     t  (p)    PV1Avg:1 PV1 SD:1
-----------------------------------------------------------------------------
  G       0.00         7       0.70   -0.24    -7.83(.000) -7.20     1.40
  H       1.00       101      10.08   -0.45   -16.15(.000) -3.07     1.43
  I       2.00       369      36.83   -0.41   -14.32(.000) -1.05     1.27
  J       3.00       373      37.23    0.26     8.65(.000)  0.91     1.44
  K       4.00       117      11.68    0.46    16.39(.000)  2.85     1.34
  L       5.00        35       3.49    0.44    15.45(.000)  4.78     1.40
```

Figure 6 Extract from the Item Analysis for the Multifaceted Analysis

The `itanal` statement's output concludes with a set of summary statistics (Figure 7). For the mean, standard deviation, variance and standard error of the mean, the scores have been scaled up so that they are reported on a scale consistent with students responding to all of the generalised items.

```
-------------------------------------------------------------------------------
In this analysis 87.51%  of the data are missing.

The following results are scaled to assume that a single response
was provided for each item.

N                              8296
Mean                          78.86
Standard Deviation            24.06
Variance                     578.92
Skewness                       0.20
Kurtosis                       0.54
Standard error of mean         0.26
===============================================================================
```

Figure 7 Summary Statistics for the Multifaceted Analysis

Note:	***Traditional methods are not well suited to multifaceted measurement. If more than 10% of the response data is missing—either at random or by design (as will often be the case in multifaceted designs)—the test reliability and standard error of measurement will not be computed.***

THE MULTIFACETED ANALYSIS RESTRICTED TO ONE CRITERION

In analysing these data with the multifaceted model, the fit statistics have suggested a lack of independence between the raters' judgments for the two criteria and evidence of unmodelled noise in the raters' behaviour. Here, therefore, an additional analysis is undertaken that adds some support to the hypothesis that the raters' OP and TF judgments are not independent. In this second analysis, only one criterion (OP) is analysed.

The files that we use in this sample analysis are:

`ex3b.cqc`	The command statements.
`ex3.dat`	The data.
`ex3b.shw`	The results of the single-criterion multifaceted analysis.

(The last file is created when the command file is executed.)

The command file for fitting the multifaceted model to these data but using only one of the criteria is given in Figure 8. The code listed here is very similar to that given in Figure 1, so we will only discuss the differences.

```
1.  Title Rater Effects Model Two;
2.  datafile ex3.dat;
3.  format rater 17-18 rater 19-20
       responses 21 responses 25 ! criteria(1);
4.  codes G,H,I,J,K,L;
5.  score (G,H,I,J,K,L) (0,1,2,3,4,5);
6.  labels 1 OP !criteria;
7.  /*labels 2 TF !criteria;*/
8.  model rater + criteria + step;
9.  Estimate ! nodes=30;
10. show ! estimates=latent >> ex3b.shw;
```

Figure 8 Command File for Fitting the Multifaceted Model to One of the Writing Criteria Only

1.-2. As in Figure 1.

3. The response blocks in the `format` statement now refer to one column only, the column that contains the OP criteria for each rater. Note that in the option we now indicate that there is just one criterion in each response block.

4.-6. As in Figure 1.

7. The `labels` statement for the TF criterion is now unnecessary, so we have enclosed it inside comment markers (/* and */).

8.-10. As for lines 8, 9, and 10 in Figure 1, except the `show` statement output is directed to a different file, `ex3b.shw`.

RUNNING THE MULTIFACETED MODEL FOR ONE CRITERION

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

```
ex3b.cqc
```

Select Run -> Run All.

ConQuest will begin executing the statements that are in the file `ex3b.cqc`; and as they are executed, they will be echoed on the screen. When ConQuest reaches the `estimate` statement, it will begin fitting the multifaceted model to the data; and as it does so, it will report on the progress of the estimation. Due to the large size of this data file, ConQuest will take some time to perform this analysis, which will take 69 iterations to converge.

In Figures 9 and 10, the rater and step parameter estimates are given for this model from the second table in the file `ex3b.shw`. The part of the table that reports on the `criteria` facet is not shown here, since there is only one criterion and it must therefore have an estimate of zero. In fact, the inclusion of the `criteria` term in the `model` statement was redundant.

```
================================================================================
Rater Effects Model Two                                   Tue Oct 03 17:05 2006
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
================================================================================
TERM 1: rater
--------------------------------------------------------------------------------
   VARIABLES                      UNWEIGHTED FIT            WEIGHTED FIT
------------                   ----------------------   ----------------------
    rater    ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------
 1   14        0.770   0.039    0.89 ( 0.92, 1.08) -2.7   0.89 ( 0.91, 1.09) -2.4
 2   17        0.070   0.039    0.97 ( 0.91, 1.09) -0.6   0.99 ( 0.91, 1.09) -0.3
 3   18       -0.039   0.041    1.50 ( 0.90, 1.10)  8.7   1.48 ( 0.90, 1.10)  8.1
 4   19       -1.320   0.037    1.03 ( 0.91, 1.09)  0.8   1.03 ( 0.92, 1.08)  0.7
 5   24        0.737   0.039    1.22 ( 0.91, 1.09)  4.7   1.22 ( 0.91, 1.09)  4.4
 6   38        0.209   0.041    0.86 ( 0.91, 1.09) -3.0   0.90 ( 0.90, 1.10) -2.0
 7   67        0.466   0.038    0.99 ( 0.92, 1.08) -0.3   1.01 ( 0.91, 1.09)  0.3
 8   70       -0.095   0.039    1.00 ( 0.91, 1.09)  0.1   0.99 ( 0.91, 1.09) -0.1
 9   73       -0.254   0.038    0.93 ( 0.92, 1.08) -1.7   0.89 ( 0.91, 1.09) -2.5
 10  74       -0.136   0.036    1.17 ( 0.92, 1.08)  4.3   1.13 ( 0.92, 1.08)  3.1
 11  78       -0.341   0.038    0.87 ( 0.91, 1.09) -3.1   0.91 ( 0.91, 1.09) -2.0
 12  79        0.124   0.038    0.83 ( 0.92, 1.08) -4.2   0.88 ( 0.91, 1.09) -2.9
 13                                                .9   0.95 ( 0.91, 1.09) -1.2
 14                                                .8   1.04 ( 0.92, 1.08)  1.0
 15                                                .7   0.92 ( 0.90, 1.10) -1.6
 16  93       -0.291^  0.150    0.94 ( 0.91, 1.09) -1.3   0.98 ( 0.91, 1.09) -0.4
--------------------------------------------------------------------------------
================================================================================
```

The fit statistics for this model are better than the corresponding fit statistics for the previous model.

Figure 9 Rater Severity Parameter Estimates

```
============
TERM 3: step
--------------------------------------------------------------------------------
   VARIABLES                      UNWEIGHTED FIT            WEIGHTED FIT
------------                   ----------------------   ----------------------
     step   ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------
       0                       0.36 ( 0.97, 1.03)-55.3   1.45 ( 0.83, 1.17)  4.6
       1     -6.007   0.056    0.96 ( 0.97, 1.03) -2.4   1.03 ( 0.95, 1.05)  1.1
       2     -3.124   0.029    1.04 ( 0.97, 1.03)  2.6   1.02 ( 0.97, 1.03)  1.7
       3      0.766   0.020    1.03 ( 0.97, 1.03)  1.9   1.04 ( 0.97, 1.03)  2.6
       4      3.170   0.031    1.08 ( 0.97, 1.03)  5.0   1.02 ( 0.95, 1.05)  1.0
       5      5.195*           0.87 ( 0.97, 1.03) -8.6   1.28 ( 0.88, 1.12)  4.3
--------------------------------------------------------------------------------
================================================================================
```

The fit statistics for this model are better than the corresponding fit statistics for the previous model.

Figure 10 Step Parameter Estimates

A comparison of Figures 9 and 10 with Figures 2, 3, and 4 shows that this second model leads to an improved fit for both the `rater` and `step` parameters. It would appear that the apparent noisy behaviour of the raters, as illustrated in Figure 2, is a result of the redundancy in the two criteria and is not evident if a single criterion is analysed. The fit statistics for the steps are similarly improved, suggesting either that the redundancy between the criteria was influencing the step fits or that there is a `rater` by `criteria` interaction.

WARNING: *t is not appropriate to use the deviance statistic to compare the fit of the two models fitted in this tutorial. The deviance statistic can only be used when one model is a submodel of the other. For this to occur, the models must result in response patterns that are the same length, and each of the items must have the same number of response categories in each of the analyses (which was not the case here).*

The dependency possibility can be further explored by using the model that assumed independence (the first sample analysis in this tutorial) to calculate the expected frequencies of various pairs of OP and TF ratings and then comparing the expected frequencies with the observed frequencies of those pairs. Figure 11 shows a two-dimensional frequency plot of the observed and expected number of scores for pairs of values of TF and OP given by rater 85. The diagonal line shows the points where the TF and OP scores are equal. It is noted that the observed frequencies are much higher than the expected frequencies along this diagonal, indicating that rater 85 tends to give more identical scores for TF and OP than one would expect. Similar patterns are also observed for other raters. It appears that a model that takes account of the severity of the rater and the difficulty of the criteria does not fit these data well.

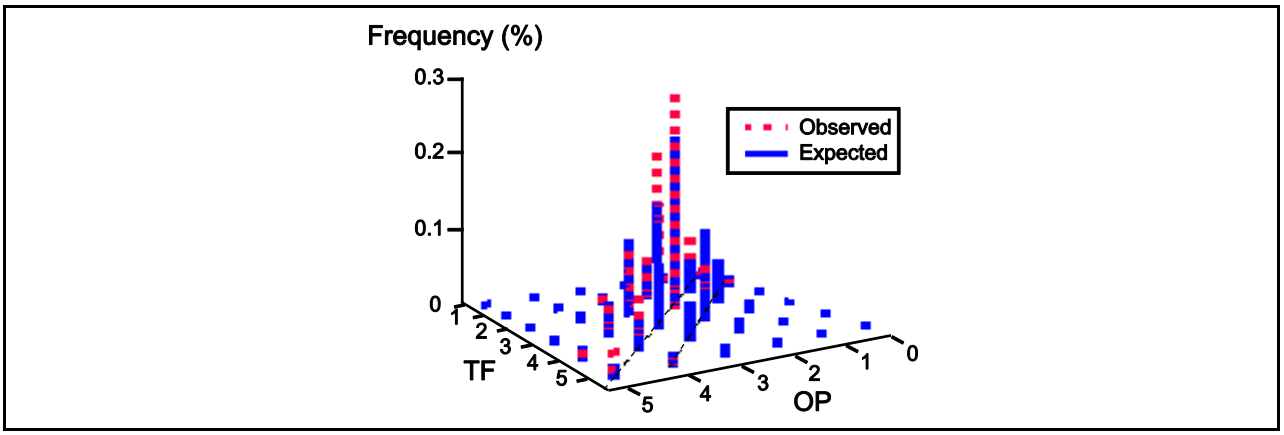


Figure 11 Observed versus Expected Frequencies for Pairs of OP and TF scores

SUMMARY

In this tutorial, we have seen how to fit multifaceted models with ConQuest. Our sample analysis has used only one additional facet (rater), but ConQuest can analyse up to 1000 facets.

Some key points we have covered in this tutorial are:

- ConQuest can be used to fit multifaceted item response models easily.
- The `format` statement is very flexible and can deal with many of the alternative ways that multifaceted data can be formatted (see the command reference for more examples).
- A `score` statement can be used to assign scores to the response categories that are modelled.
- We have reiterated the point that response categories and item scores are *not* the same thing.
- Fit statistics can be used to suggest alternative models that might be fitted to the data.

REFERENCES

Congdon, P., and. McQueen, J. 1997. The stability of rater severity estimates in large scale performance assessment programmes. Paper presented at the Annual Meeting of the American Educational Research Association. March 24–28, Chicago, Illinois.

Fischer, G. H. 1973. The linear logistic model as an instrument in educational research. *Acta Psychologica, 37*, 359–74.

Linacre, J. M. 1994. *Many-Facet Rasch Measurement*. Chicago. MESA Press (original work published 1989).