Many Facets and Hierarchical Model Testing

Ray Adams and Margaret Wu, 27 August 2010

In tutorial three, the notion of additional measurement facets is introduced, and data was analysed with one additional facet, a rater facet. The number of facets that can be used with multifaceted measurement models is theoretically unlimited, although, as shall be seen in this tutorial, the addition of each new facet adds considerably to the range of models that need to be considered.[1] A number of techniques are available for choosing between alternative models for multifaceted data. First, the deviance statistic of alternative models can be compared to provide a formal statistical test of the relative fit of models. Second, the fit statistics for the parameter estimates can be used, as was done in tutorial three. Third, the estimated values of the parameters associated with a term in a model can be examined to see if that term is necessary. In this tutorial, we illustrate these strategies for choosing between the many alternative multifaceted models that can be applied to data that have more than two facets.

FITTING A GENERAL THREE-FACETED MODEL

The data that we are analysing in this tutorial are simulated three-faceted data.[2] The data were simulated to reflect an assessment context in which 500 students have each provided written responses to two out of a total of four writing topics. Each of these tasks was then rated by two out of four raters against five assessment criteria. For each of the five criteria, a four-point rating scale was used with codes 0, 1, 2 and 3. This results in four sets of ratings (two essay topics by two raters' judgments) against the five criteria for each of the 500 students. In generating the data, two raters and two topics were randomly assigned to the students, and the model used assumed that the raters differed in harshness, that the criteria differed in difficulty, and that the rating structure varied across the criteria. The topics were assumed to be of equal difficulty; there were no interactions between the `topic`, `criteria` and `rater` facets; and the step structure did not vary with `rater` or `topic`.

In the first analysis, we fit a model that assumes main effects for all facets, the set of three two-way interactions, and a step structure that varies with `topic`, `item` and `rater`. The files used in this sample analysis are:

`ex4a.cqc`	The command statements that used for the first analysis.
`ex4.dat`	The data.
`ex4.nam`	The variable labels for the facet elements.
`ex4a.prm`	Initial values for the item parameter estimates.
`ex4a.reg`	Initial values for the regression parameter estimates.
`ex4a.cov`	Initial values for the variance parameter estimates.
`ex4a.shw`	Selected results of the first analysis.
`ex4b.cqc`	The command statements used for the second analysis.

[1] ConQuest can model up to 1000 different facets.

[2] For those familiar with Linacre's ([1989] 1994) approach and terminology, these would be considered four-faceted data, since Linacre counts the cases as a facet, whereas we count the unique variables in the `model` statement.

`ex4b_1.shw` and `ex4b_2.shw`	Selected results of the second analysis.
`ex4c.cqc`	The command statements used for the third analysis.
`ex4c_1.shw` `ex4c_11.shw`	Selected results of the third analysis.

(The `.prm`, `.reg`, `.cov`, and `.shw` files are created when the command file is executed.)

The data were entered into the file `ex4.dat` using four lines per student, one for each rater and topic combination. For each of the lines, column 1 contains a rater code, column 3 contains a topic code and columns 5 through 9 contain the ratings of the five criteria given by the matching rater and topic combination. The command file for fitting one possible multifaceted model to these data is shown in Figure 1.

```
1.   datafile ex4.dat;
2.   format     rater 1 topic 3 responses 5-9 /
                rater 1 topic 3 responses 5-9 /
                rater 1 topic 3 responses 5-9 /
                rater 1 topic 3 responses 5-9 !
                criteria(5);
3.   labels << ex4.nam;
4.   set update=yes,warnings=no;
5.   model  rater + topic + criteria + rater*topic +
            rater*criteria + topic*criteria +
            rater*topic*criteria*step;
6.   export parameters >> ex4a.prm;
7.   export reg_coefficients >> ex4a.reg;
8.   export covariance >> ex4a.cov;
9.   estimate!nodes=10,stderr=full;
10.  show   parameters ! estimates=latent, tables=1:2:4 >> ex4a.shw;
```

Figure 1 Sample Command File for a Very General Multifaceted Model

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

2. Multifaceted data can be entered into data sets in many ways. The ConQuest `format` statement is very flexible and can cater for many alternative data specifications. Here the data are spread over four lines for each student. Each line contains a rater code, a topic code and five responses. The slash (/) character is used to indicate that the following data should be read from the next line of the data file. The multiple use of the terms `rater`, `topic` and `responses` allows us to read the multiple sets of ratings for each student. In this case, the term `rater` is used four times, `topic` four times and `responses` four times. Thus, the rater and topic indicated on the first line for each case will be associated with the responses on the first line, the rater and topic on the second line will be associated with the responses on the second line, and so on. More generally, if variables are repeated in a `format` statement, the *n*-th occurrence of `responses` will be associated with the *n*-th occurrence of any other variable, or the *n*-th occurrence of `responses` will be matched with the highest occurrence of any other variable if *n* is greater than the number of occurrences of that variable.

   This `format` statement also includes an option, `criteria(5)`, which assigns the variable name `criteria` to the five responses that are implicitly identified by the response block. If this option had been omitted, the default variable name for the responses would have been `item`.

3. The labels for the facets in this analysis are to be read from the file `ex4.nam`. The contents of this file are shown in Figure 2. Here we have provided labels for each of the three facets. The character string `===>` precedes the name of the facet, and the following lines contain the facet level and then the label that is to be assigned to that level.

```
===> rater
1  Amy
2  Beverely
3  Colin
4  David
===> topic
1  Sport
2  Family
3  Work
4  School
===> criteria
1  spelling
2  coherence
3  structure
4  grammar
5  content
```

Figure 2 The Labels File for the Many Facets Sample Analysis

4. The `set` statement can be used to alter some of ConQuest's default values. In this case, the default status of the `update` and `warnings` settings has been changed. When `update` is set to `yes`, in conjunction with the following `export` statements, updated parameter estimates will be written to a file at the completion of every iteration. This option is particularly valuable when analyses take a long time to execute. If the `update` option is set to `yes` and you have to terminate the analysis for some reason (e.g., you want to use the computer for something else and ConQuest is monopolising CPU time), you can interrupt the job and then restart it at some later stage with starting values set to the most recent parameter estimates. (To use these starting values, you would have to add one or more `import` statements to the command file.)

   Setting `warnings` to `no` tells ConQuest not to report warning messages. Errors, however, will still be reported. Setting `warnings` to `no` is typically used in conjunction with setting `update` to `yes` in order to suppress the warning message that there is a file overwrite at every iteration.

5. The `model` statement contains seven terms: `rater`, `topic`, `criteria`, `rater*topic`, `rater*criteria`, `topic*criteria`, and `rater*topic*criteria*step`. This `model` statement indicates that seven sets of parameters are to be estimated. The first three are main effects and correspond to a set of rater harshness parameters, a set of topic difficulty parameters, and a set of criteria difficulty parameters. The next three are two-way interactions between the facets. The first of these interaction terms models a variation in rater harshness across the topics (or, equivalently, variation in topic difficulty across the raters), the second models a variation in rater harshness across the criteria, and the third represents a variation in the topic difficulties across the criteria. The final term represents a set of parameters to describe the step structure of the responses. The step structure is modelled as varying across all combinations of raters, topics and criteria.

   One additional term could be added to this model: the three-way interaction between raters, topics and criteria.

6.-8. The `export` statements request that the parameter estimates be written to text files in a simple, unlabelled format. The `export` statement can be used to produce files that are more readily read by other software. Further, the format of each export file matches the format of ConQuest import files so that export files that are written by ConQuest can be re-read as either anchor files or initial value files.

9. The `estimate` statement initiates the estimation of the item response model. In this case, two options are used to change the default settings of the estimation procedures. The `nodes=10` option means that the numerical integration that is necessary in the estimation will be done with a Gauss-hermite quadrature method using 10 nodes.[3] The default number of nodes is 15, but we have chosen to reduce the number of nodes to 10 for this sample analysis, since it will reduce the processing time. Simulation results by Wu and Adams (1993) illustrate that 10 nodes will normally be sufficient for accurate estimation. The `stderr=full` option causes ConQuest to compute the full error variance-covariance matrix for the model that has been estimated. This method provides the most accurate estimates of the asymptotic error variances that ConQuest can compute. It does, however, take a considerable amount of computing time, even on very fast machines. In 'Estimating Standard Errors' in Chapter 12 of Wu, Adams, Wilson and Haldane (2007), we discuss the circumstances under which it is desirable to use the `stderr=full` option. In this case, we have used it because of the large number of facets, each of which has only a couple of levels.

10. The `show` statement produces a display of the item response model parameter estimates and saves them to the file `ex4a.shw`. The option `estimates=latent` requests that the displays include an illustration of the latent ability distribution. The option `tables=1:2:4` limits the output to tables 1, 2 and 4.

RUNNING THE MULTIFACETED SAMPLE ANALYSIS

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

```
Ex4a.cqc
```

Select Run -> Run All. ConQuest will begin executing the statements that are in the file ex4a.cqc; and as they are executed, they will be echoed in the Output window. 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 this estimation. This analysis will take 703 iterations to converge, and the calculation of the standard errors may take a considerable amount of time.

After the estimation is complete, the output from the `show` statement can be found in the file `ex4a.shw`. Figures 3 and 4 are extracts from the second table in this file.

Figure 3 shows the parameter estimates for the three main effects: rater, topic and criteria. Notice that the separation reliability for the topic is close to zero and that the variation between the topic parameter estimates is not significant. This result suggests that the `topic` term might be deleted from the model because the topics do not vary in their difficulty. (Thus, ConQuest has confirmed the model we used in our data simulation.)

[3] See 'Estimation' in Chapter 12 of Wu *et al.* (2007) for further explanation of the estimation methods that are used in ConQuest.

```
============================================================================
ConQuest: Generalised Item Response Modelling Software     Wed Oct 04 12:23 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   Amy             -0.871   0.042    1.00 ( 0.81, 1.19) -0.0   0.99 ( 0.74, 1.26) -0.1
 2   Beverely        -0.537   0.035    1.09 ( 0.82, 1.18)  1.0   1.03 ( 0.78, 1.22)  0.3
 3   Colin            0.452   0.030    0.98 ( 0.81, 1.19) -0.2   0.98 ( 0.80, 1.20) -0.2
 4   David            0.956*          1.09 ( 0.81, 1.19)  0.9   1.08 ( 0.79, 1.21)  0.7
----------------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.997
Chi-square test of parameter equality =     894.87,  df = 3,  Sig Level = 0.000
============================================================================
TERM 2: topic
----------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT            WEIGHTED FIT
---------------                     ------------------------   ------------------------
    topic           ESTIMATE  ERROR    MNSQ       CI         T    MNSQ       CI         T
----------------------------------------------------------------------------------------
 1   Sport           -0.023   0.031    1.00 ( 0.81, 1.19) -0.0   0.97 ( 0.79, 1.21) -0.3
 2   Family           0.016   0.033    1.09 ( 0.81, 1.19)  0.9   1.08 ( 0.79, 1.21)  0.7
 3   Work             0.005   0.031    1.00 ( 0.81, 1.19)  0.1   0.99 ( 0.78, 1.22) -0.1
 4   School           0.002*          0.98 ( 0.81, 1.19) -0.1   0.99 ( 0.79, 1.21) -0.1
----------------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.000
Chi-square test of parameter equality =       0.82,  df = 3,  Sig Level = 0.845
============================================================================
TERM 3: criteria
----------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT            WEIGHTED FIT
---------------                     ------------------------   ------------------------
   criteria         ESTIMATE  ERROR    MNSQ       CI         T    MNSQ       CI         T
----------------------------------------------------------------------------------------
 1   spelling        -1.046   0.048    1.01 ( 0.88, 1.12)  0.2   0.98 ( 0.84, 1.16) -0.3
 2   coherence       -0.569   0.037    1.03 ( 0.88, 1.12)  0.6   1.08 ( 0.84, 1.16)  1.0
 3   structure       -0.051   0.035    0.96 ( 0.88, 1.12) -0.6   0.93 ( 0.86, 1.14) -1.0
 4   grammar          0.551   0.029    1.09 ( 0.88, 1.12)  1.4   1.09 ( 0.86, 1.14)  1.2
 5   content          1.116*          1.05 ( 0.88, 1.12)  0.8   1.07 ( 0.87, 1.13)  1.1
----------------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.997
Chi-square test of parameter equality =    1078.20,  df = 4,  Sig Level = 0.000
```

Figure 3 The Parameter Estimates for Rater Severity, Topic Difficulty and Criterion Difficulty

Figure 4 shows the parameter estimates for one of the three two-way interaction terms. The results reported in this figure suggest that there is no interaction between the topic and criterion. (Again, ConQuest has confirmed the model we used in our data simulation.) The results for the two remaining two-way interaction terms are not reported here; however, if you examine them in the file `ex4a.shw` you will see that, although the effects are statistically significant, they are very small and we could probably ignore them.

```
============================================================================
TERM 6: topic*criteria
-------------------------------------------------------------------------------------------------------
          VARIABLES                          UNWEIGHTED FIT            WEIGHTED FIT
--------------------------                 ----------------------   ----------------------
    topic      criteria    ESTIMATE  ERROR   MNSQ       CI        T    MNSQ       CI        T
-------------------------------------------------------------------------------------------------------
 1   Sport   1   spelling     0.057   0.069   0.87 ( 0.81, 1.19) -1.4   0.92 ( 0.74, 1.26) -0.6
 2   Family  1   spelling    -0.031   0.074   0.90 ( 0.81, 1.19) -1.0   0.98 ( 0.75, 1.25) -0.1
 3   Work    1   spelling    -0.091   0.073   1.08 ( 0.81, 1.19)  0.8   0.95 ( 0.73, 1.27) -0.3
 4   School  1   spelling     0.065*          0.95 ( 0.81, 1.19) -0.5   1.03 ( 0.74, 1.26)  0.3
 1   Sport   2   coherence   -0.045   0.055   1.18 ( 0.81, 1.19)  1.9   1.19 ( 0.74, 1.26)  1.4
 2   Family  2   coherence    0.050   0.057   1.13 ( 0.81, 1.19)  1.3   1.10 ( 0.74, 1.26)  0.8
 3   Work    2   coherence    0.003   0.053   1.05 ( 0.81, 1.19)  0.6   1.04 ( 0.74, 1.26)  0.3
 4   School  2   coherence   -0.008*          0.80 ( 0.81, 1.19) -2.2   0.93 ( 0.74, 1.26) -0.6
 1   Sport   3   structure    0.014   0.051   0.93 ( 0.81, 1.19) -0.7   0.99 ( 0.79, 1.21) -0.1
 2   Family  3   structure   -0.018   0.054   1.03 ( 0.81, 1.19)  0.4   0.99 ( 0.78, 1.22) -0.1
 3   Work    3   structure    0.015   0.051   1.08 ( 0.81, 1.19)  0.8   1.03 ( 0.77, 1.23)  0.3
 4   School  3   structure   -0.012*          0.95 ( 0.81, 1.19) -0.4   0.88 ( 0.78, 1.22) -1.2
 1   Sport   4   grammar     -0.029   0.047   1.07 ( 0.81, 1.19)  0.7   1.07 ( 0.79, 1.21)  0.6
 2   Family  4   grammar     -0.016   0.050   1.15 ( 0.81, 1.19)  1.5   1.08 ( 0.78, 1.22)  0.7
 3   Work    4   grammar      0.050   0.048   0.95 ( 0.81, 1.19) -0.5   0.97 ( 0.78, 1.22) -0.2
 4   School  4   grammar     -0.004*          1.12 ( 0.81, 1.19)  1.3   1.16 ( 0.79, 1.21)  1.4
 1   Sport   5   content      0.002*          1.02 ( 0.81, 1.19)  0.2   0.96 ( 0.80, 1.20) -0.3
 2   Family  5   content      0.015*          0.96 ( 0.81, 1.19) -0.3   1.02 ( 0.79, 1.21)  0.2
 3   Work    5   content      0.023*          1.15 ( 0.81, 1.19)  1.5   1.15 ( 0.79, 1.21)  1.4
 4   School  5   content     -0.041*          0.89 ( 0.81, 1.19) -1.2   0.91 ( 0.80, 1.20) -0.9
-------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.000
Chi-square test of parameter equality =       5.66,  df = 12,  Sig Level = 0.932
============================================================================
```

Figure 4 Parameter Estimates for the `topic*criteria` Interaction

THE FIT OF TWO ADDITIONAL ALTERNATIVE MODELS

Many submodels of the model analysed with the command file in Figure 1 can be fitted to these data. As we mentioned above, the model that was actually used in the generation of these data can be fitted by replacing the `model` statement in Figure 1 with `model rater + criteria + criteria*step`. The file `ex4b.cqc` contains statements that will fit this submodel and an even simpler model (`rater + step`). The item response model parameter estimates that are obtained from the first of these models are shown in Figure 5. As would be expected, the fit for each of the parameters is good.

The other important thing to note about Figure 5 is the values of the parameter estimates. When the data in `ex4.dat` were generated, the rater parameters were set at –1.0, –0.5, 0.5 and 1.0 and the criteria parameters were set at –1.2, –0.6, 0, 0.6 and 1.2.

Figure 6 shows the item parameter estimates when the `model` statement is changed to `model rater + step`, which assumes that there is no variation between the criteria in difficulty, a simplification that we know does not hold for these data. The fact that this model is not appropriate for the data can be easily identified by the fact that the deviance has increased significantly from the deviance for the model that was fit in Figure 5 (as shown in the first

```
================================================================================
ConQuest: Generalised Item Response Modelling Software     Wed Oct 04 15:00 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   Amy            -0.999   0.029    1.01 ( 0.81, 1.19)  0.2   0.98 ( 0.75, 1.25) -0.1
 2   Beverely       -0.550   0.025    1.05 ( 0.82, 1.18)  0.6   1.01 ( 0.78, 1.22)  0.1
 3   Colin           0.518   0.024    0.98 ( 0.81, 1.19) -0.2   0.99 ( 0.80, 1.20) -0.0
 4   David           1.032*           1.05 ( 0.81, 1.19)  0.5   1.03 ( 0.79, 1.21)  0.3
--------------------------------------------------------------------------------
An asterisk next to a parameter esti                        it is constrained
Separation Reliability =  0.999
Chi-square test of parameter equalit                = 3,          0.000
================================================================================
TERM 2: criteria
--------------------------------------------------------------------------------
   VARIABLES                          UNWEIGHTED FIT          WEIGHTED FIT
---------------                   ----------------------   ----------------------
   criteria        ESTIMATE  ERROR    MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------
 1   spelling       -1.192   0.039    1.07 ( 0.88, 1.12)  1.1   1.01 ( 0.84, 1.16)  0.2
 2   coherence      -0.591   0.031    1.07 ( 0.88, 1.12)  1.1   1.08 ( 0.84, 1.16)  1.0
 3   structure       0.007   0.028    0.94 ( 0.88, 1.12) -1.0   0.92 ( 0.86, 1.14) -1.2
 4   grammar         0.617   0.026    1.07 ( 0.88, 1.12)  1.1   1.06 ( 0.86, 1.14)  0.9
 5   content         1.158*           1.03 ( 0.88, 1.12)  0.5   1.05 ( 0.87, 1.13)  0.8
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.998
Chi-square test of parameter equality =    1865.48,  df = 4,  Sig Level = 0.000
================================================================================
TERM 3: criteria*step
--------------------------------------------------------------------------------
          VARIABLES                          UNWEIGHTED FIT          WEIGHTED FIT
--------------------              ----------------------   ----------------------
   criteria    step  ESTIMATE  ERROR    MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------
 1   spelling    0                      0.44 ( 0.88, 1.12)-11.2   0.87 ( 0.72, 1.28) -0.9
 1   spelling    1   -0.362   0.116    1.01 ( 0.88, 1.12)  0.2   1.01 ( 0.79, 1.21)  0.1
 1   spelling    2   -0.226   0.110    1.07 ( 0.88, 1.12)  1.1   1.03 ( 0.87, 1.13)  0.4
 1   spelling    3    0.588*          1.08 ( 0.88, 1.12)  1.3   1.05 ( 0.87, 1.13)  0.8
 2   coherence   0                      1.66 ( 0.88, 1.12)  8.7   1.08 ( 0.81, 1.19)  0.8
 2   coherence   1    0.614   0.107    0.75 ( 0.88, 1.12) -4.2   0.90 ( 0.79, 1.21) -1.0
 2   coherence   2   -0.303   0.118    0.97 ( 0.88, 1.12) -0.5   0.99 ( 0.85, 1.15) -0.1
 2   coherence   3   -0.311*          1.00 ( 0.88, 1.12)  0.0   1.06 ( 0.86, 1.14)  0.9
 3   structure   0                      1.01 ( 0.88, 1.12)  0.1   0.95 ( 0.84, 1.16) -0.5
 3   structure   1   -0.198   0.075    0.81 ( 0.88, 1.12) -3.1   0.90 ( 0.85, 1.15) -1.3
 3   structure   2   -0.163   0.082    0.90 ( 0.88, 1.12) -1.7   0.92 ( 0.87, 1.13) -1.3
 3   structure   3    0.361*          0.89 ( 0.88, 1.12) -1.7   0.92 ( 0.87, 1.13) -1.3
 4   grammar     0                      1.52 ( 0.88, 1.12)  7.1   1.07 ( 0.86, 1.14)  1.0
 4   grammar     1    0.116   0.069    0.91 ( 0.88, 1.12) -1.5   0.92 ( 0.86, 1.14) -1.1
 4   grammar     2    0.104   0.086    1.00 ( 0.88, 1.12)  0.0   0.99 ( 0.86, 1.14) -0.1
 4   grammar     3   -0.220*          0.99 ( 0.88, 1.12) -0.2   0.97 ( 0.87, 1.13) -0.4
 5   content     0                      1.15 ( 0.88, 1.12)  2.3   1.07 ( 0.87, 1.13)  1.1
 5   content     1   -0.314   0.060    1.02 ( 0.88, 1.12)  0.3   1.02 ( 0.87, 1.13)  0.3
 5   content     2    0.077   0.077    1.02 ( 0.88, 1.12)  0.3   1.05 ( 0.86, 1.14)  0.7
 5   content     3    0.237*          0.94 ( 0.88, 1.12) -0.9   0.99 ( 0.86, 1.14) -0.1
--------------------------------------------------------------------------------
```

Parameter estimates are close to their generating values.

Fit statistics look good.

Figure 5 Parameter Estimates for `model rater + criteria + criteria*step;`

table generated by the `show` statement). This observation is discussed in detail in the next section ('A Sequence of Models'). From Figure 6, however, we note that the fit statistics, at least in the case of the rater parameters, are smaller than they should be. When lower than expected fit statistic values are found, it is generally a result of unmodelled dependencies in the data. In the tutorial three, we saw that low fit was probably due to an unmodelled dependency between the two criteria, OP and TF. Here the low fit suggests that there is an unmodelled consistency between the rater judgments. The judgments across raters are more consistent than the model expects, and this has arisen because an element of consistency between judgments in the ratings can be traced to the variance in the criteria difficulties, a variation that is not currently being modelled.

```
================================================================================
ConQuest: Generalised Item Response Modelling Software     Wed Oct 04 15:00 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   Amy          -0.669   0.022    0.84 ( 0.81, 1.19) -1.7   0.87 ( 0.74, 1.26) -1.0
 2   Beverely     -0.322   0.019    0.90 ( 0.82, 1.18) -1.1   0.84 ( 0.78, 1.22) -1.5
 3   Colin         0.333   0.018    0.88 ( 0.81, 1.19) -1.3   0.86 ( 0.79, 1.21) -1.4
 4   David         0.658*           0.93 ( 0.81, 1.19) -0.7   0.94 ( 0.79, 1.21) -0.6
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.998
Chi-square test of parameter equality =    1518.90,  df = 3
================================================================
TERM 2: step
--------------------------------------------------------------------------------
   VARIABLES                          UNWEIGHTED FIT             WEIGHTED FIT
--------------                     ----------------------   ----------------------
     step        ESTIMATE  ERROR    MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------
       0                            1.01 ( 0.88, 1.12)  0.2   0.99 ( 0.85, 1.15) -0.2
       1           0.415   0.033    0.94 ( 0.88, 1.12) -1.0   0.92 ( 0.86, 1.14) -1.1
       2          -0.046   0.039    1.01 ( 0.88, 1.12)  0.2   1.03 ( 0.87, 1.13)  0.5
       3          -0.369*           1.02 ( 0.88, 1.12)  0.3   1.05 ( 0.87, 1.13)  0.7
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
```

These fit statistics are all negative.

Figure 6 Parameter Estimates for `model rater + step;`

A SEQUENCE OF MODELS

A search for a model that provides the most parsimonious fit to these data can be undertaken in a systematic fashion by using hierarchical model fitting techniques in conjunction with the use of the chi-squared test of parameter equality. The file `ex4c.cqc` includes 11 ConQuest runs, the results of which are written to the files `ex4c_1.shw` through `ex4c_11.shw`. Figure 7 illustrates the hierarchy of models that are included in `ex4c.cqc` and summarises the fit of the models. Notice, as we move through the hierarchy from model (1) to model (5) and then model (9), how the fit is not significantly worsened by removing terms. The same is also true if we follow the path (1) to (3) and then (7) to (9). Comparing models (5) and (6), we note that the `rater` term is necessary—that is, there is significant variation between the raters in their harshness. Comparing models (9) and (10), we can see that the step parameters vary significantly with the criteria.

SUMMARY

In this tutorial, we have seen how ConQuest can be used to compare the fit of competing models that may be considered appropriate for a data set. We have seen how to use the deviance statistics, fit statistics and test of parameter equality to assist in the choice of a best fitting model.

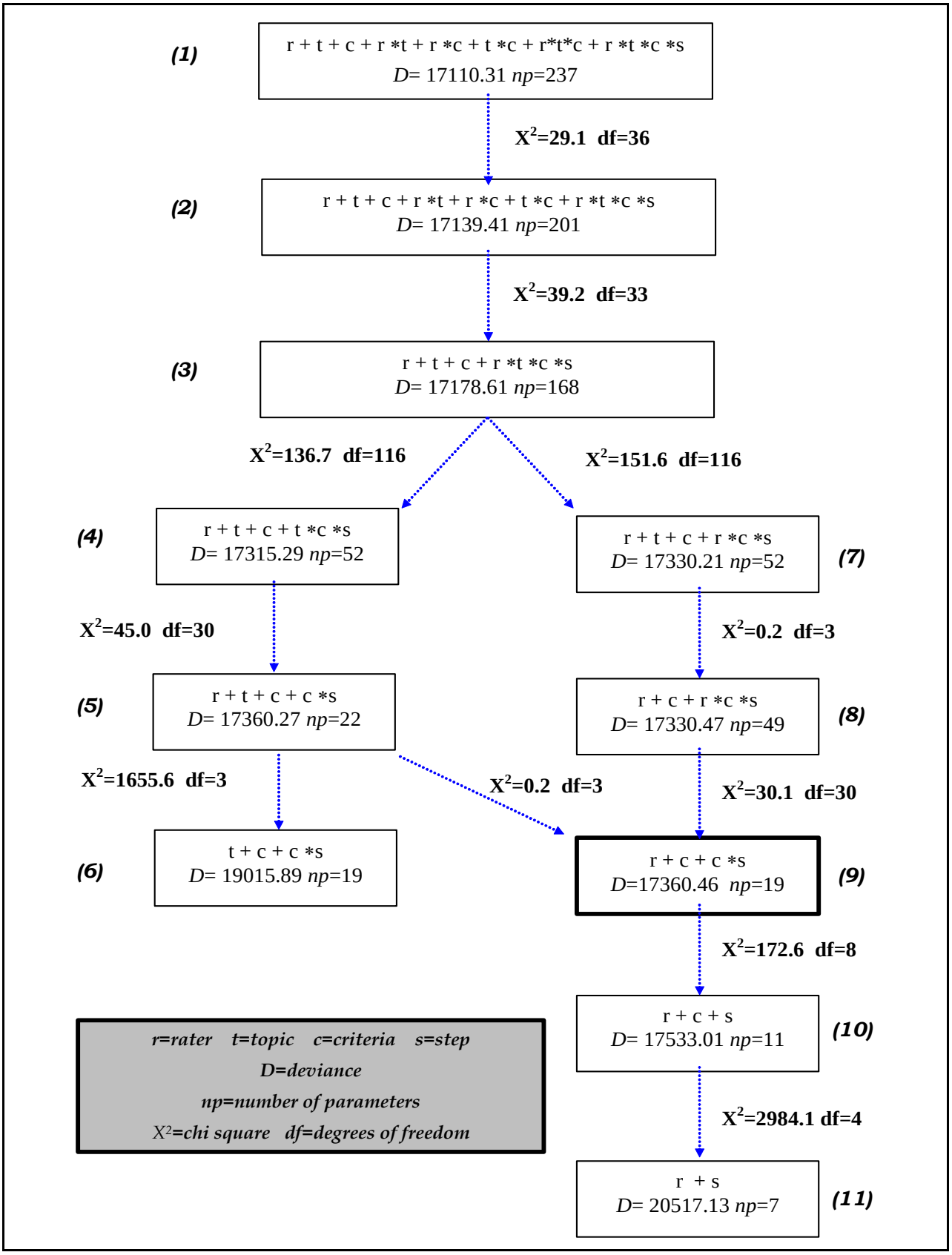

Figure 7 A Hierarchy of Models and Their Fit

REFERENCES

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

Wu, M. L., and Adams, R. J. 1993. Simulating parameter recovery for the random coefficients multinomial logit. Paper presented at the Fifth International Objective Measurement Workshop. April, Atlanta, Georgia.

Wu, M. L., Adams, R. J., Wilson, M. R., Haldane, S.A. (2007). *ACER ConQuest Version 2: Generalised item response modelling software* [computer program]. Camberwell: Australian Council for Educational Research.