MONASH College

UECA Assessment Symposium 2019

The use of many-facet Rasch analysis in improving rater reliability in rating writing assessments

Phuong Tran

Monash University English Language Centre

DEFINITIONS OF TERMS

- **Rater**: a person that rates ratees' responses to constructed response tasks
- **Scoring rubrics**: a set of criteria with descriptors for different levels of performance
- **Criterion score:** a rating for a criterion
- **Task score**: the total score of all ratings for a ratee's performance on a task

MANY FACETS IN PERFORMANCE ASSESSMENT

What factors can affect students' scores on a constructed response task?

- **Ratings** assigned to responses do NOT depend only on items & tasks:
  - **Item/task difficulty**
  - **Student ability**
- **Other facets may affect ratings** (e.g., raters and rating criteria)
  - **Rater consistency / reliability + Rater severity / leniency**
  - **Rating criteria goodness of fit**

DEFINITIONS OF TERMS

- **Rater consistency**

  A tendency of a rater to assign **the same scores** to papers of the same performance levels (at both criterion level and task level)

- **Rater severity**

  A tendency of a rater to assign **scores that on average are lower than expected** if the scores given by other raters to the same group of test takers are taken into consideration.

- **Rater leniency**

  A tendency of a rater to assign **scores that on average are higher than expected** if the scores given by other raters to the same group of test takers are taken into consideration

RATING BEHAVIOURS

How do you know if a rater is consistent and appropriate in rating?

RATER ANALYSIS USING MANY-FACET RASCH MEASUREMENT (MFRM)

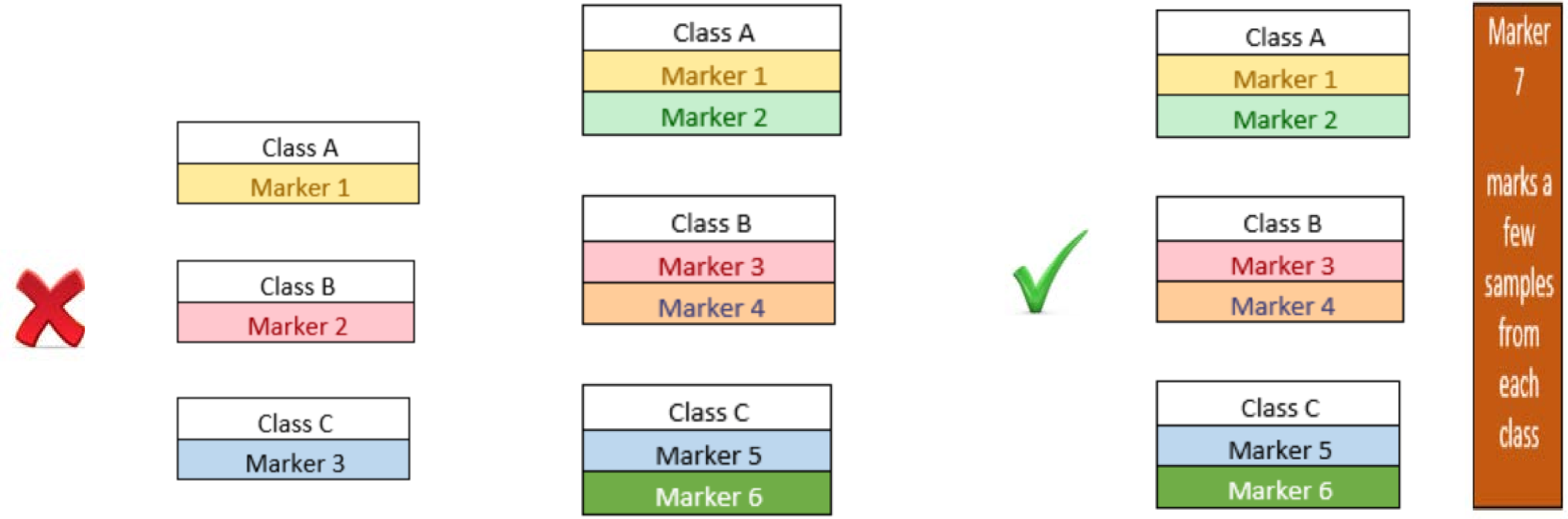
- **MFRM models** are **mathematical models** constructed to explain the relationship among facets. It performs the **logistic transformation** of ratios of successive category probabilities.
- **Independent variables**: test takers, raters, task, criteria
- **Dependent variables**: probability of getting a score category
- **Raters** are analysed based on their ratings to all the students they rate.
- **Raters** are analysed in relation to one another.

RATER ANALYSIS USING MANY-FACET RASCH MEASUREMENT

- **The MFRM** simultaneously and independently analyses the impact of different facets and calibrates the impact into one common log-linear scale (logit scale).
- Students' ability levels are controlled for, so **ratings** can be fairly evaluated.
- Rater severity is controlled for, so **examinee measures** can be calculated (i.e., independent of the variation in rater severity).
- It gives **a fair measure** of the students' performance – measures that would be obtained if raters were equally lenient/harsh.

DATA COLLECTION

- Papers are at least double-marked.
- **Raters** need at least **50 score points** (**13 papers** x 4 criterion scores) for stable estimation of rater measures (Linacre, 1994).
- Raters are **linked** via **common papers**.



✗

Class A
Marker 1

Class B
Marker 2

Class C
Marker 3

Single marked

Class A
Marker 1
Marker 2

Class B
Marker 3
Marker 4

Class C
Marker 5
Marker 6

Double marked

✓

Class A
Marker 1
Marker 2

Class B
Marker 3
Marker 4

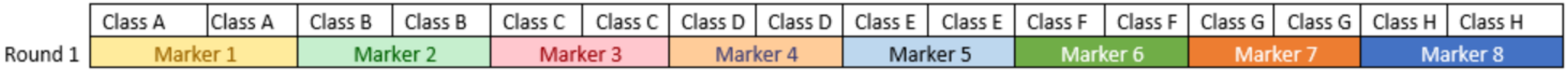
Class C
Marker 5
Marker 6

Marker 7 marks a few samples from each class

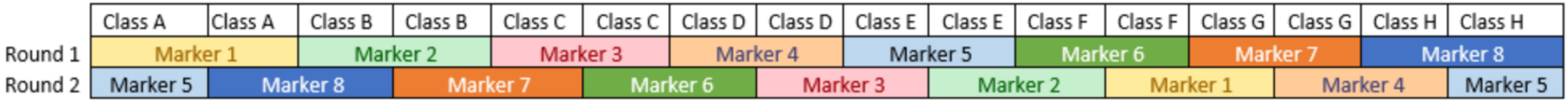
Double marked, rater linked

DATA COLLECTION

- Example of another way to link papers:



	Class A	Class A	Class B	Class B	Class C	Class C	Class D	Class D	Class E	Class E	Class F	Class F	Class G	Class G	Class H	Class H
Round 1	Marker 1		Marker 2		Marker 3		Marker 4		Marker 5		Marker 6		Marker 7		Marker 8	

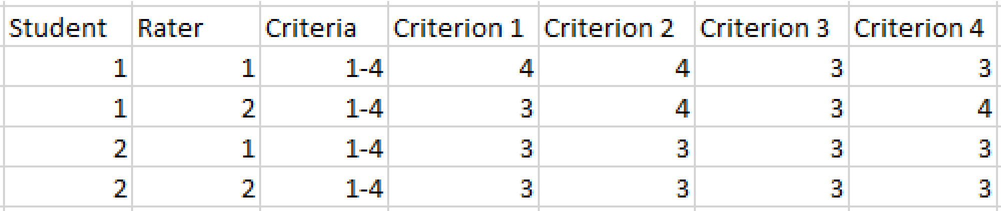


	Class A	Class A	Class B	Class B	Class C	Class C	Class D	Class D	Class E	Class E	Class F	Class F	Class G	Class G	Class H	Class H
Round 1	Marker 1		Marker 2		Marker 3		Marker 4		Marker 5		Marker 6		Marker 7		Marker 8	
Round 2	Marker 5	Marker 8		Marker 7		Marker 6		Marker 3		Marker 2		Marker 1		Marker 4		Marker 5

- Papers can be distributed among raters in many different ways, as long as **raters are linked** to one another by **each pair marking a few common papers**.

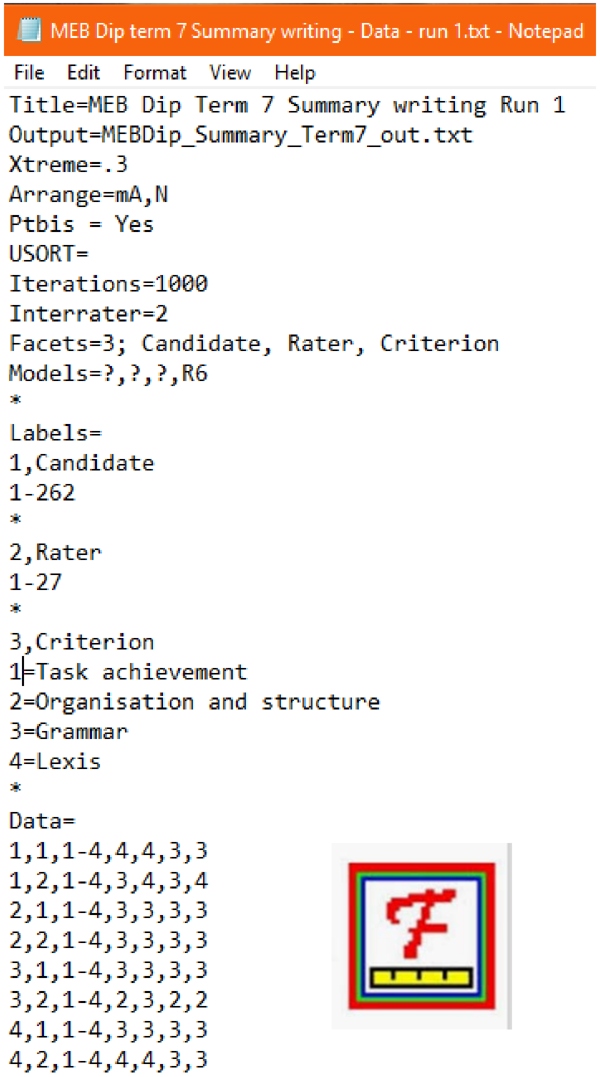
DATA ENTRY, MANIPULATION, AND ANALYSIS

- **Criterion scores** recorded for **all raters** and **all candidates**.
- **Students** are **coded** (if necessary).
- **Raters** are **coded** (if necessary).



Student	Rater	Criteria	Criterion 1	Criterion 2	Criterion 3	Criterion 4
1	1	1-4	4	4	3	3
1	2	1-4	3	4	3	4
2	1	1-4	3	3	3	3
2	2	1-4	3	3	3	3

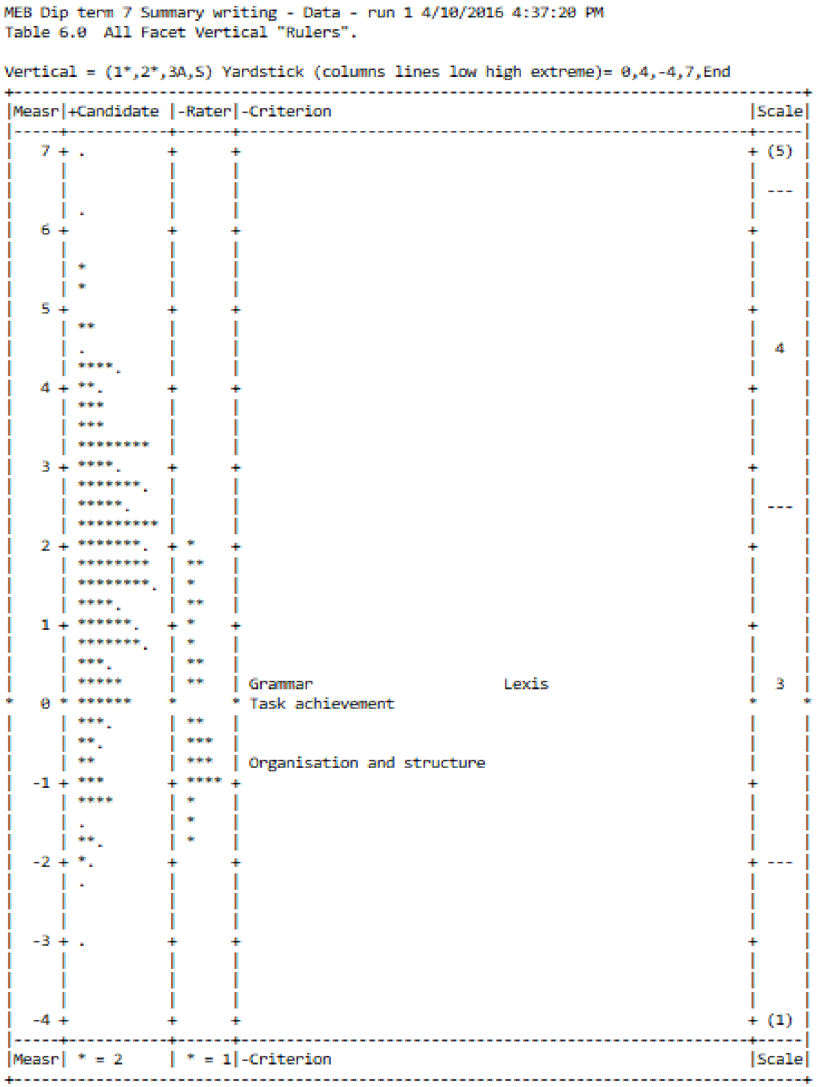
- **Control file** are written with specifications of the model.
- Data is analysed using **Facets** (Linacre, 2015).



MEB Dip term 7 Summary writing - Data - run 1.txt - Notepad
File Edit Format View Help

```
Title=MEB Dip Term 7 Summary writing Run 1
Output=MEBDip_Summary_Term7_out.txt
Xtreme=.3
Arrange=mA,N
Ptbis = Yes
USORT=
Iterations=1000
Interrater=2
Facets=3; Candidate, Rater, Criterion
Models=?,?,?,R6
*
Labels=
1,Candidate
1-262
*
2,Rater
1-27
*
3,Criterion
1=Task achievement
2=Organisation and structure
3=Grammar
4=Lexis
*
Data=
1,1,1-4,4,4,3,3
1,2,1-4,3,4,3,4
2,1,1-4,3,3,3,3
2,2,1-4,3,3,3,3
3,1,1-4,3,3,3,3
3,2,1-4,2,3,2,2
4,1,1-4,3,3,3,3
4,2,1-4,4,4,3,3
```

OUTPUT – ALL FACET VERTICAL ‘RULERS’

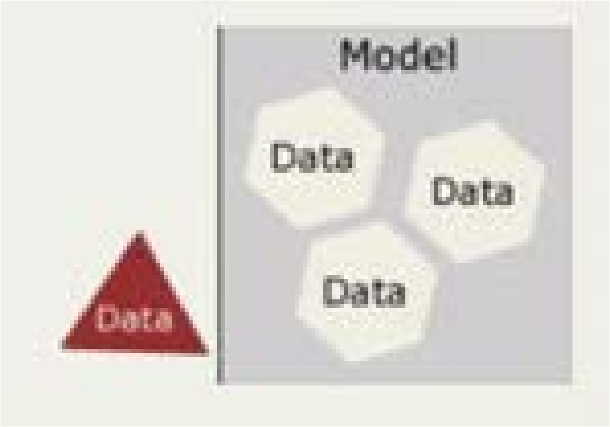


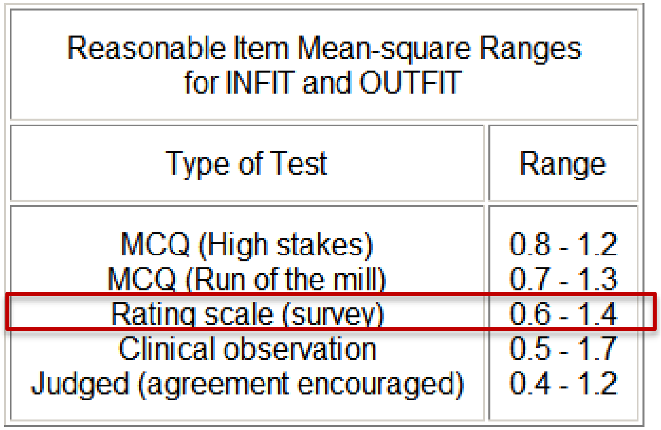
```
MEB Dip term 7 Summary writing - Data - run 1 4/10/2016 4:37:20 PM
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1*,2*,3A,S) Yardstick (columns lines low high extreme)= 0,4,-4,7,End
+--------------------------------------------------------------------------------------+
|Measr|+Candidate |-Rater|-Criterion                                             |Scale|
|-----+-----------+------+-------------------------------------------------------+-----|
|   7 + .         +      +                                                       + (5) |
|     |           |      |                                                       |     |
|     |           |      |                                                       | --- |
|     | .         |      |                                                       |     |
|   6 +           +      +                                                       +     |
|     |           |      |                                                       |     |
|     | *         |      |                                                       |     |
|     | *         |      |                                                       |     |
|   5 +           +      +                                                       +     |
|     | **        |      |                                                       |     |
|     | .         |      |                                                       |  4  |
|     | ****.     |      |                                                       |     |
|   4 + **.       +      +                                                       +     |
|     | ***       |      |                                                       |     |
|     | ***       |      |                                                       |     |
|     | ********  |      |                                                       |     |
|   3 + ****.     +      +                                                       +     |
|     | *******.  |      |                                                       |     |
|     | *****.    |      |                                                       | --- |
|     | ********* |      |                                                       |     |
|   2 + *******.  + *    +                                                       +     |
|     | ********  | **   |                                                       |     |
|     | ********. | *    |                                                       |     |
|     | ****.     | **   |                                                       |     |
|   1 + ******.   + *    +                                                       +     |
|     | *******.  | *    |                                                       |     |
|     | ***.      | **   |                                                       |     |
|     | *****     | **   | Grammar                    Lexis                      |  3  |
*   0 * ******    *      * Task achievement                                      *     *
|     | ***.      | **   |                                                       |     |
|     | **.       | ***  |                                                       |     |
|     | **        | ***  | Organisation and structure                            |     |
|  -1 + ***       + **** +                                                       +     |
|     | ****      | *    |                                                       |     |
|     | .         | *    |                                                       |     |
|     | **.       | *    |                                                       |     |
|  -2 + *.        +      +                                                       + --- |
|     | .         |      |                                                       |     |
|     |           |      |                                                       |     |
|     |           |      |                                                       |     |
|  -3 + .         +      +                                                       +     |
|     |           |      |                                                       |     |
|     |           |      |                                                       |     |
|     |           |      |                                                       |     |
|  -4 +           +      +                                                       + (1) |
|-----+-----------+------+-------------------------------------------------------+-----|
|Measr| * = 2     | * = 1|-Criterion                                             |Scale|
+--------------------------------------------------------------------------------------+
```

SOME EVALUATION CRITERIA

- **Degree of consistency: Goodness of fit**





Reasonable Item Mean-square Ranges for INFIT and OUTFIT

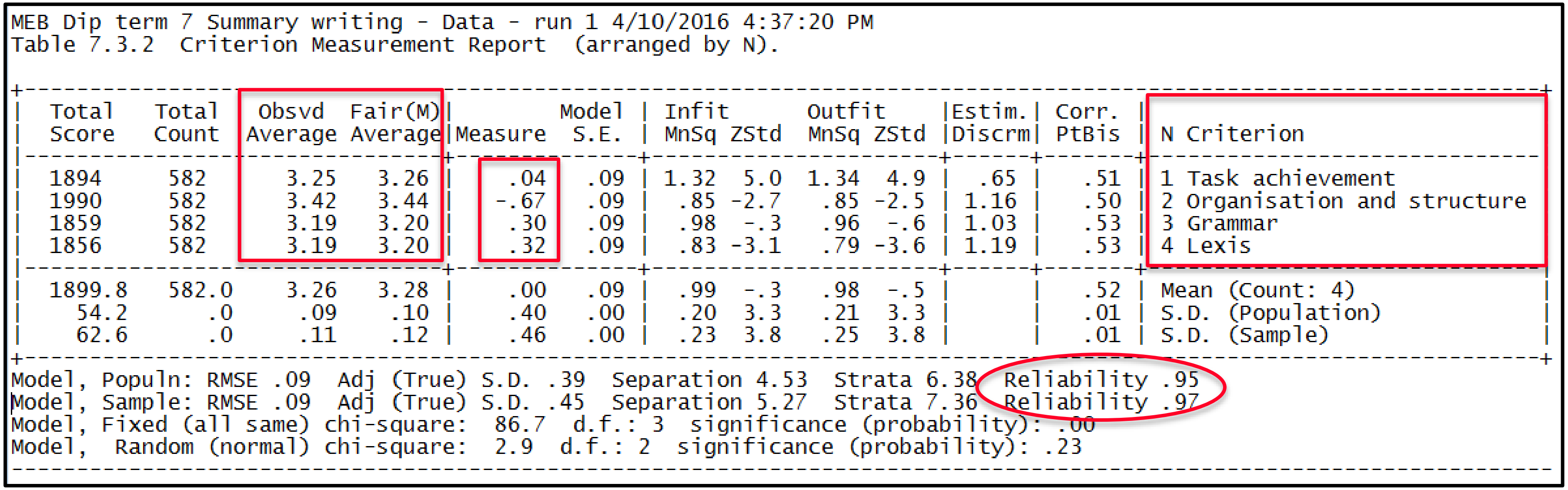
Type of Test	Range
MCQ (High stakes)	0.8 - 1.2
MCQ (Run of the mill)	0.7 - 1.3
Rating scale (survey)	0.6 - 1.4
Clinical observation	0.5 - 1.7
Judged (agreement encouraged)	0.4 - 1.2

- **Reliability** (Rater separation index): As close to **0** as possible => higher rater agreement
- **Degree of appropriateness:**

Commonly agreed: Measure/Standard error: ≥2.0 (harsh) or ≤-2.0 (lenient)

Definitely target: Measure/Standard error: ≥5.0 (harsh) or ≤-5.0 (lenient)

OUTPUT – CRITERION MEASUREMENT REPORT

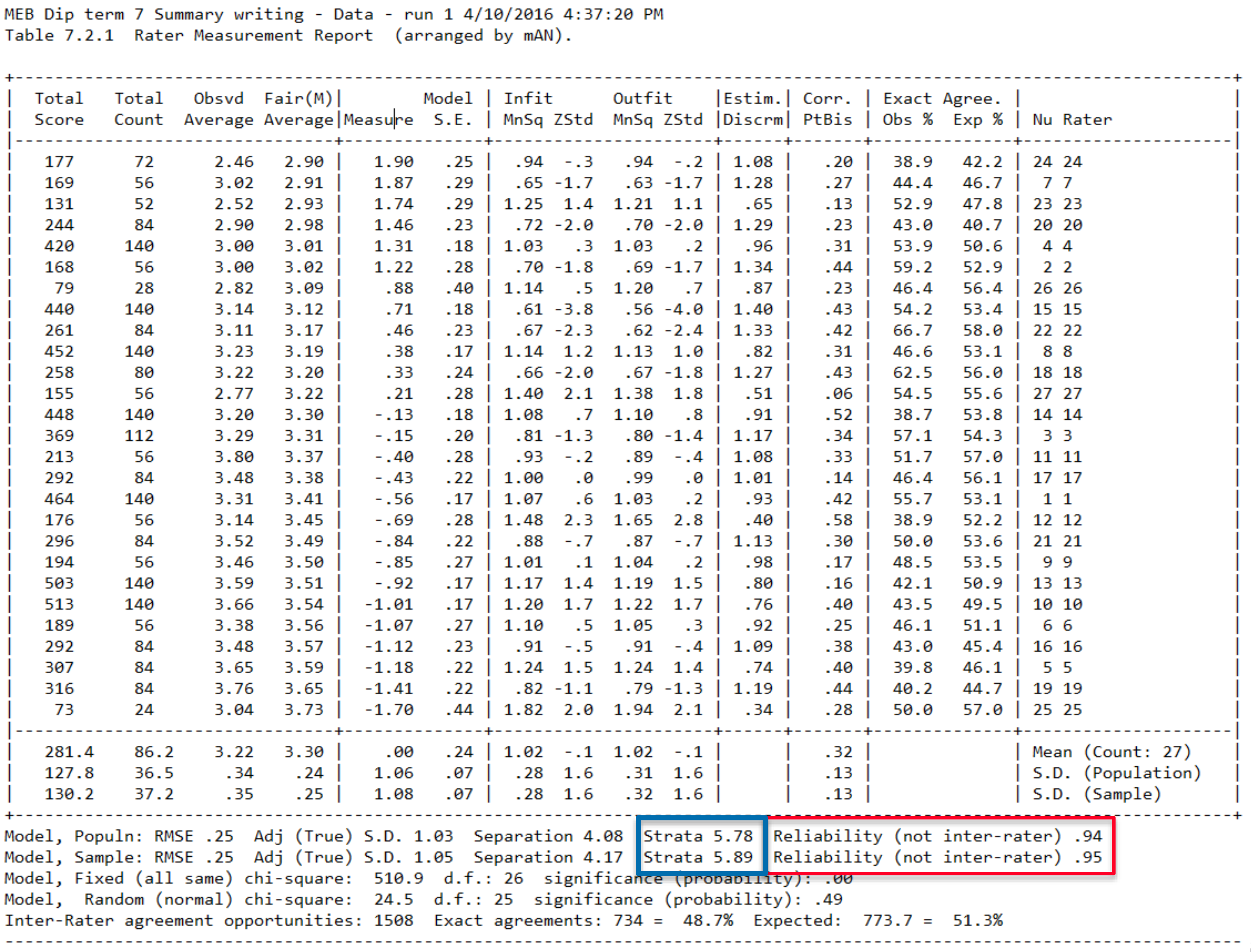


```
MEB Dip term 7 Summary writing - Data - run 1 4/10/2016 4:37:20 PM
Table 7.3.2  Criterion Measurement Report  (arranged by N).

+------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|       Model | Infit      Outfit    |Estim.| Corr. |                        |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtBis | N Criterion            |
|--------------------------------+--------------+----------------------+------+-------+------------------------|
|  1894     582     3.25   3.26 |    .04   .09 | 1.32  5.0  1.34  4.9 |  .65 |   .51 | 1 Task achievement     |
|  1990     582     3.42   3.44 |   -.67   .09 |  .85 -2.7   .85 -2.5 | 1.16 |   .50 | 2 Organisation and structure
|  1859     582     3.19   3.20 |    .30   .09 |  .98  -.3   .96  -.6 | 1.03 |   .53 | 3 Grammar              |
|  1856     582     3.19   3.20 |    .32   .09 |  .83 -3.1   .79 -3.6 | 1.19 |   .53 | 4 Lexis                |
|--------------------------------+--------------+----------------------+------+-------+------------------------|
|  1899.8   582.0   3.26   3.28 |    .00   .09 |  .99  -.3   .98  -.5 |      |   .52 | Mean (Count: 4)        |
|    54.2      .0    .09    .10 |    .40   .00 |  .20  3.3   .21  3.3 |      |   .01 | S.D. (Population)      |
|    62.6      .0    .11    .12 |    .46   .00 |  .23  3.8   .25  3.8 |      |   .01 | S.D. (Sample)          |
+------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .09  Adj (True) S.D. .39  Separation 4.53  Strata 6.38  Reliability .95
Model, Sample: RMSE .09  Adj (True) S.D. .45  Separation 5.27  Strata 7.36  Reliability .97
Model, Fixed (all same) chi-square:  86.7  d.f.: 3  significance (probability): .00
Model,  Random (normal) chi-square:  2.9  d.f.: 2  significance (probability): .23
--------------------------------------------------------------------------------------------------------------
```

4 2 1 3

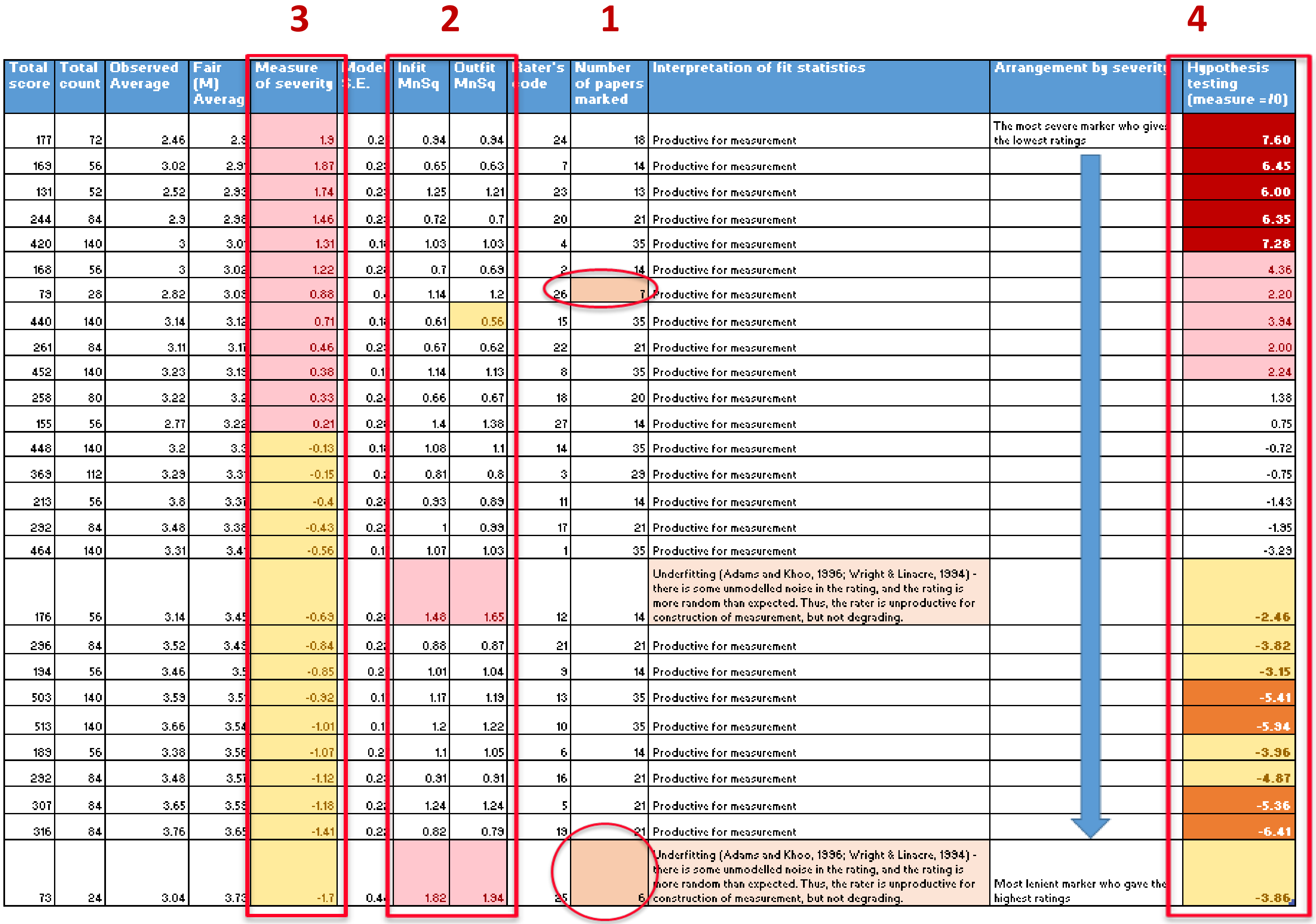
OUTPUT – RATER MEASUREMENT REPORT



```
MEB Dip term 7 Summary writing - Data - run 1 4/10/2016 4:37:20 PM
Table 7.2.1  Rater Measurement Report  (arranged by mAN).

+-----------------------------------------------------------------------------------------------------------------------+
|  Total    Total   Obsvd  Fair(M)|          Model | Infit       Outfit     |Estim.| Corr.  | Exact Agree. |              |
|  Score    Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtBis  | Obs %   Exp % | Nu Rater     |
|---------------------------------+--------------+----------------------+------+--------+--------------+--------------|
|   177      72     2.46    2.90  |  1.90    .25 |  .94  -.3    .94  -.2 | 1.08 |   .20  |  38.9   42.2 | 24 24        |
|   169      56     3.02    2.91  |  1.87    .29 |  .65 -1.7    .63 -1.7 | 1.28 |   .27  |  44.4   46.7 |  7 7         |
|   131      52     2.52    2.93  |  1.74    .29 | 1.25  1.4  1.21   1.1 |  .65 |   .13  |  52.9   47.8 | 23 23        |
|   244      84     2.90    2.98  |  1.46    .23 |  .72 -2.0    .70 -2.0 | 1.29 |   .23  |  43.0   40.7 | 20 20        |
|   420     140     3.00    3.01  |  1.31    .18 | 1.03   .3  1.03    .2 |  .96 |   .31  |  53.9   50.6 |  4 4         |
|   168      56     3.00    3.02  |  1.22    .28 |  .70 -1.8    .69 -1.7 | 1.34 |   .44  |  59.2   52.9 |  2 2         |
|    79      28     2.82    3.09  |   .88    .40 | 1.14   .5  1.20    .7 |  .87 |   .23  |  46.4   56.4 | 26 26        |
|   440     140     3.14    3.12  |   .71    .18 |  .61 -3.8    .56 -4.0 | 1.40 |   .43  |  54.2   53.4 | 15 15        |
|   261      84     3.11    3.17  |   .46    .23 |  .67 -2.3    .62 -2.4 | 1.33 |   .42  |  66.7   58.0 | 22 22        |
|   452     140     3.23    3.19  |   .38    .17 | 1.14  1.2  1.13   1.0 |  .82 |   .31  |  46.6   53.1 |  8 8         |
|   258      80     3.22    3.20  |   .33    .24 |  .66 -2.0    .67 -1.8 | 1.27 |   .43  |  62.5   56.0 | 18 18        |
|   155      56     2.77    3.22  |   .21    .28 | 1.40  2.1  1.38   1.8 |  .51 |   .06  |  54.5   55.6 | 27 27        |
|   448     140     3.20    3.30  |  -.13    .18 | 1.08   .7  1.10    .8 |  .91 |   .52  |  38.7   53.8 | 14 14        |
|   369     112     3.29    3.31  |  -.15    .20 |  .81 -1.3    .80 -1.4 | 1.17 |   .34  |  57.1   54.3 |  3 3         |
|   213      56     3.80    3.37  |  -.40    .28 |  .93  -.2    .89  -.4 | 1.08 |   .33  |  51.7   57.0 | 11 11        |
|   292      84     3.48    3.38  |  -.43    .22 | 1.00   .0    .99   .0 | 1.01 |   .14  |  46.4   56.1 | 17 17        |
|   464     140     3.31    3.41  |  -.56    .17 | 1.07   .6  1.03    .2 |  .93 |   .42  |  55.7   53.1 |  1 1         |
|   176      56     3.14    3.45  |  -.69    .28 | 1.48  2.3  1.65   2.8 |  .40 |   .58  |  38.9   52.2 | 12 12        |
|   296      84     3.52    3.49  |  -.84    .22 |  .88  -.7    .87  -.7 | 1.13 |   .30  |  50.0   53.6 | 21 21        |
|   194      56     3.46    3.50  |  -.85    .27 | 1.01   .1  1.04    .2 |  .98 |   .17  |  48.5   53.5 |  9 9         |
|   503     140     3.59    3.51  |  -.92    .17 | 1.17  1.4  1.19   1.5 |  .80 |   .16  |  42.1   50.9 | 13 13        |
|   513     140     3.66    3.54  | -1.01    .17 | 1.20  1.7  1.22   1.7 |  .76 |   .40  |  43.5   49.5 | 10 10        |
|   189      56     3.38    3.56  | -1.07    .27 | 1.10   .5  1.05    .3 |  .92 |   .25  |  46.1   51.1 |  6 6         |
|   292      84     3.48    3.57  | -1.12    .23 |  .91  -.5    .91  -.4 | 1.09 |   .38  |  43.0   45.4 | 16 16        |
|   307      84     3.65    3.59  | -1.18    .22 | 1.24  1.5  1.24   1.4 |  .74 |   .40  |  39.8   46.1 |  5 5         |
|   316      84     3.76    3.65  | -1.41    .22 |  .82 -1.1    .79 -1.3 | 1.19 |   .44  |  40.2   44.7 | 19 19        |
|    73      24     3.04    3.73  | -1.70    .44 | 1.82  2.0  1.94   2.1 |  .34 |   .28  |  50.0   57.0 | 25 25        |
|---------------------------------+--------------+----------------------+------+--------+--------------+--------------|
|  281.4    86.2    3.22    3.30  |   .00    .24 | 1.02  -.1  1.02   -.1 |      |   .32  |              | Mean (Count: 27)     |
|  127.8    36.5     .34     .24  |  1.06    .07 |  .28  1.6   .31   1.6 |      |   .13  |              | S.D. (Population)    |
|  130.2    37.2     .35     .25  |  1.08    .07 |  .28  1.6   .32   1.6 |      |   .13  |              | S.D. (Sample)        |
+-----------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .25  Adj (True) S.D. 1.03  Separation 4.08  Strata 5.78  Reliability (not inter-rater) .94
Model, Sample: RMSE .25  Adj (True) S.D. 1.05  Separation 4.17  Strata 5.89  Reliability (not inter-rater) .95
Model, Fixed (all same) chi-square:  510.9  d.f.: 26  significance (probability): .00
Model,  Random (normal) chi-square:  24.5  d.f.: 25  significance (probability): .49
Inter-Rater agreement opportunities: 1508  Exact agreements: 734 =  48.7%  Expected:  773.7 =  51.3%
-------------------------------------------------------------------------------------------------------------------------
```

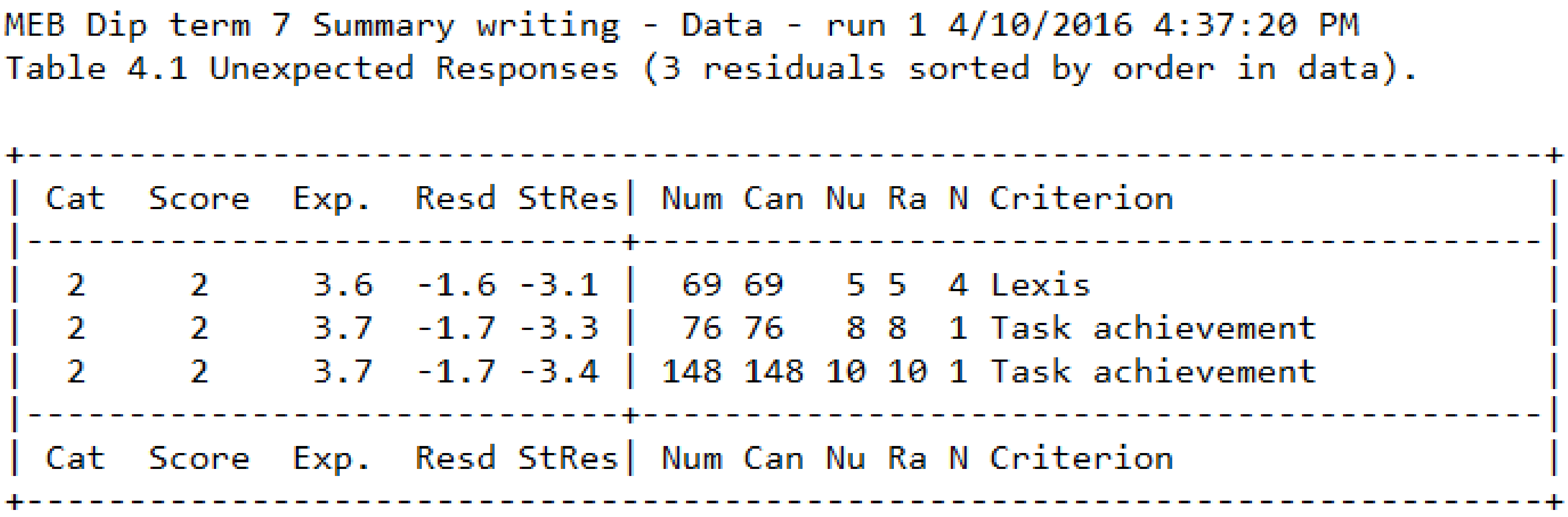
INTERPRETATION OF RESULTS



Total score	Total count	Observed Average	Fair (M) Averag	Measure of severity	Model S.E.	Infit MnSq	Outfit MnSq	Rater's code	Number of papers marked	Interpretation of fit statistics	Arrangement by severity	Hypothesis testing (measure =/0)
177	72	2.46	2.9	1.9	0.2	0.94	0.94	24	18	Productive for measurement	The most severe marker who gives the lowest ratings	7.60
169	56	3.02	2.91	1.87	0.2	0.65	0.63	7	14	Productive for measurement		6.45
131	52	2.52	2.93	1.74	0.2	1.25	1.21	23	13	Productive for measurement		6.00
244	84	2.9	2.98	1.46	0.2	0.72	0.7	20	21	Productive for measurement		6.35
420	140	3	3.01	1.31	0.1	1.03	1.03	4	35	Productive for measurement		7.28
168	56	3	3.02	1.22	0.2	0.7	0.69	2	14	Productive for measurement		4.36
79	28	2.82	3.09	0.88	0.4	1.14	1.2	26	7	Productive for measurement		2.20
440	140	3.14	3.12	0.71	0.1	0.61	0.56	15	35	Productive for measurement		3.94
261	84	3.11	3.17	0.46	0.2	0.67	0.62	22	21	Productive for measurement		2.00
452	140	3.23	3.19	0.38	0.1	1.14	1.13	8	35	Productive for measurement		2.24
258	80	3.22	3.2	0.33	0.2	0.66	0.67	18	20	Productive for measurement		1.38
155	56	2.77	3.22	0.21	0.2	1.4	1.38	27	14	Productive for measurement		0.75
448	140	3.2	3.3	-0.13	0.1	1.08	1.1	14	35	Productive for measurement		-0.72
369	112	3.29	3.31	-0.15	0.	0.81	0.8	3	29	Productive for measurement		-0.75
213	56	3.8	3.37	-0.4	0.2	0.93	0.89	11	14	Productive for measurement		-1.43
292	84	3.48	3.38	-0.43	0.2	1	0.99	17	21	Productive for measurement		-1.95
464	140	3.31	3.41	-0.56	0.1	1.07	1.03	1	35	Productive for measurement		-3.29
176	56	3.14	3.45	-0.69	0.2	1.48	1.65	12	14	Underfitting (Adams and Khoo, 1996; Wright & Linacre, 1994) - there is some unmodelled noise in the rating, and the rating is more random than expected. Thus, the rater is unproductive for construction of measurement, but not degrading.		-2.46
296	84	3.52	3.49	-0.84	0.2	0.88	0.87	21	21	Productive for measurement		-3.82
194	56	3.46	3.5	-0.85	0.2	1.01	1.04	9	14	Productive for measurement		-3.15
503	140	3.59	3.51	-0.92	0.1	1.17	1.19	13	35	Productive for measurement		-5.41
513	140	3.66	3.54	-1.01	0.1	1.2	1.22	10	35	Productive for measurement		-5.94
189	56	3.38	3.56	-1.07	0.2	1.1	1.05	6	14	Productive for measurement		-3.96
292	84	3.48	3.57	-1.12	0.2	0.91	0.91	16	21	Productive for measurement		-4.87
307	84	3.65	3.59	-1.18	0.2	1.24	1.24	5	21	Productive for measurement		-5.36
316	84	3.76	3.65	-1.41	0.2	0.82	0.79	19	21	Productive for measurement		-6.41
73	24	3.04	3.73	-1.7	0.4	1.82	1.94	25	6	Underfitting (Adams and Khoo, 1996; Wright & Linacre, 1994) - there is some unmodelled noise in the rating, and the rating is more random than expected. Thus, the rater is unproductive for construction of measurement, but not degrading.	Most lenient marker who gave the highest ratings	-3.86

MONASH College

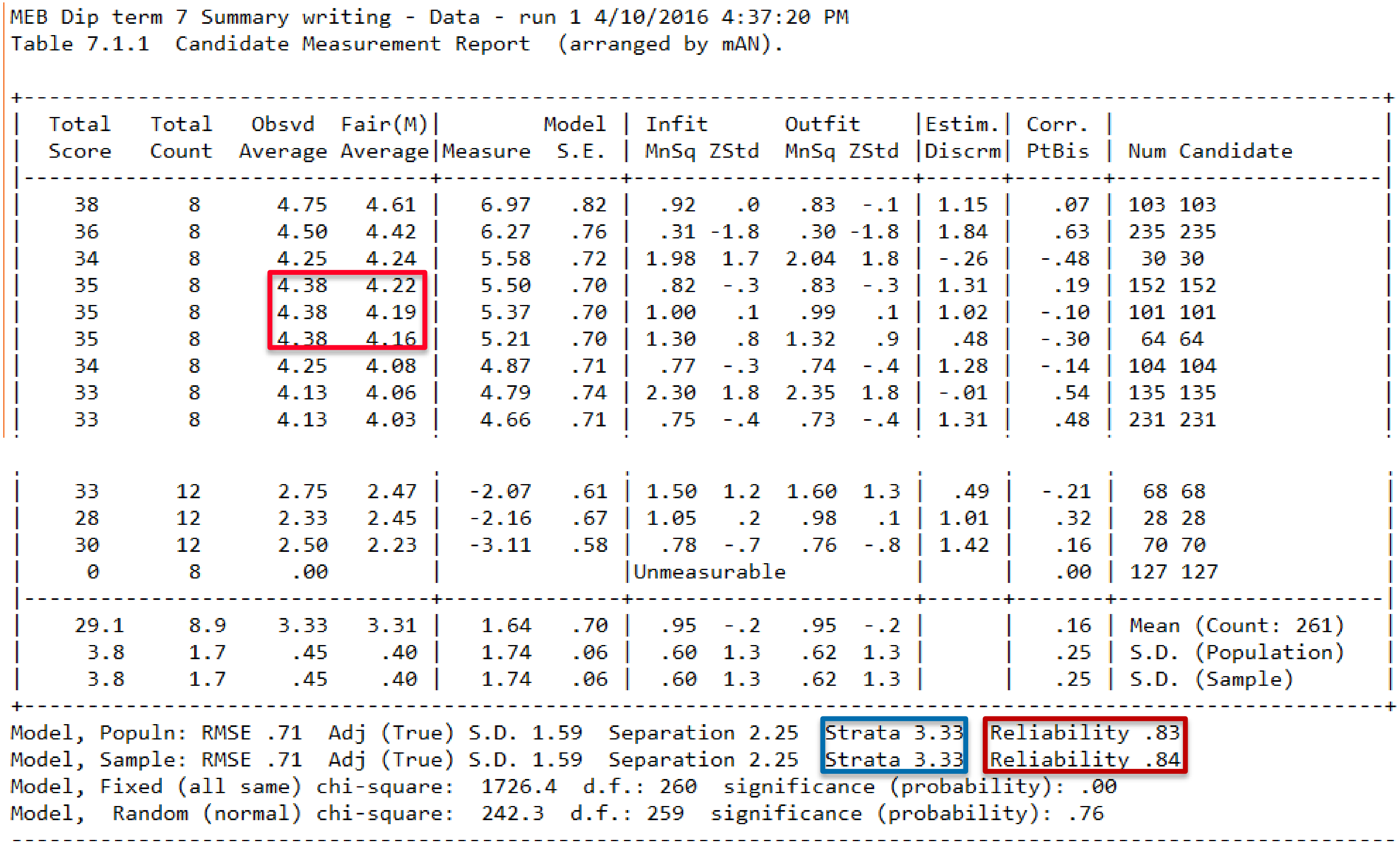
OUTPUT – UNEXPECTED RESPONSES



```
MEB Dip term 7 Summary writing - Data - run 1 4/10/2016 4:37:20 PM
Table 4.1 Unexpected Responses (3 residuals sorted by order in data).

+-------------------------------------------------------------------------------+
| Cat  Score  Exp.  Resd StRes| Num Can Nu Ra N Criterion                       |
|-----------------------------+-------------------------------------------------|
|  2     2     3.6  -1.6 -3.1 |  69 69   5 5  4 Lexis                           |
|  2     2     3.7  -1.7 -3.3 |  76 76   8 8  1 Task achievement                |
|  2     2     3.7  -1.7 -3.4 | 148 148 10 10 1 Task achievement                |
|-----------------------------+-------------------------------------------------|
| Cat  Score  Exp.  Resd StRes| Num Can Nu Ra N Criterion                       |
+-------------------------------------------------------------------------------+
```

OUTPUT – CANDIDATE MEASUREMENT REPORT

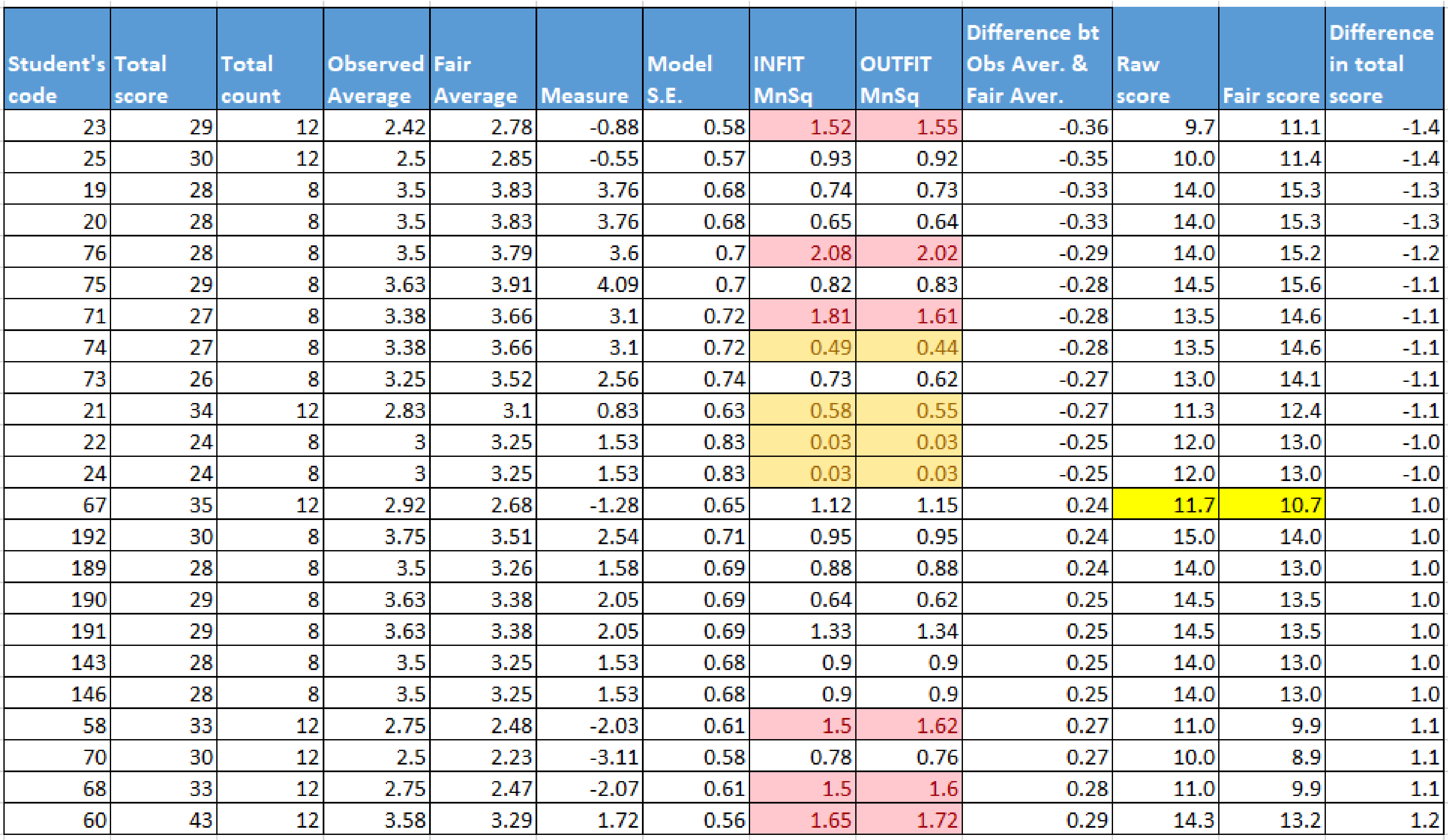


```
MEB Dip term 7 Summary writing - Data - run 1 4/10/2016 4:37:20 PM
Table 7.1.1  Candidate Measurement Report  (arranged by mAN).

+------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit       Outfit    |Estim.| Corr. |                    |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtBis | Num Candidate      |
|--------------------------------+--------------+-----------------------+------+-------+--------------------|
|    38       8     4.75   4.61  |  6.97    .82 |  .92   .0   .83  -.1 | 1.15 |   .07 | 103 103            |
|    36       8     4.50   4.42  |  6.27    .76 |  .31 -1.8   .30 -1.8 | 1.84 |   .63 | 235 235            |
|    34       8     4.25   4.24  |  5.58    .72 | 1.98  1.7  2.04  1.8 | -.26 |  -.48 |  30 30             |
|    35       8     4.38   4.22  |  5.50    .70 |  .82  -.3   .83  -.3 | 1.31 |   .19 | 152 152            |
|    35       8     4.38   4.19  |  5.37    .70 | 1.00   .1   .99   .1 | 1.02 |  -.10 | 101 101            |
|    35       8     4.38   4.16  |  5.21    .70 | 1.30   .8  1.32   .9 |  .48 |  -.30 |  64 64             |
|    34       8     4.25   4.08  |  4.87    .71 |  .77  -.3   .74  -.4 | 1.28 |  -.14 | 104 104            |
|    33       8     4.13   4.06  |  4.79    .74 | 2.30  1.8  2.35  1.8 | -.01 |   .54 | 135 135            |
|    33       8     4.13   4.03  |  4.66    .71 |  .75  -.4   .73  -.4 | 1.31 |   .48 | 231 231            |
.                                .              .                       .      .       .                    .

.                                .              .                       .      .       .                    .
|    33      12     2.75   2.47  | -2.07    .61 | 1.50  1.2  1.60  1.3 |  .49 |  -.21 |  68 68             |
|    28      12     2.33   2.45  | -2.16    .67 | 1.05   .2   .98   .1 | 1.01 |   .32 |  28 28             |
|    30      12     2.50   2.23  | -3.11    .58 |  .78  -.7   .76  -.8 | 1.42 |   .16 |  70 70             |
|     0       8      .00         |              |Unmeasurable           |      |   .00 | 127 127            |
|--------------------------------+--------------+-----------------------+------+-------+--------------------|
|    29.1     8.9   3.33   3.31  |  1.64    .70 |  .95  -.2   .95  -.2 |      |   .16 | Mean (Count: 261)  |
|     3.8     1.7    .45    .40  |  1.74    .06 |  .60  1.3   .62  1.3 |      |   .25 | S.D. (Population)  |
|     3.8     1.7    .45    .40  |  1.74    .06 |  .60  1.3   .62  1.3 |      |   .25 | S.D. (Sample)      |
+------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .71  Adj (True) S.D. 1.59  Separation 2.25  Strata 3.33  Reliability .83
Model, Sample: RMSE .71  Adj (True) S.D. 1.59  Separation 2.25  Strata 3.33  Reliability .84
Model, Fixed (all same) chi-square:  1726.4  d.f.: 260  significance (probability): .00
Model,  Random (normal) chi-square:  242.3  d.f.: 259  significance (probability): .76
--------------------------------------------------------------------------------------------------------
```

OUTPUT – CANDIDATE MEASUREMENT REPORT



Student's code	Total score	Total count	Observed Average	Fair Average	Measure	Model S.E.	INFIT MnSq	OUTFIT MnSq	Difference bt Obs Aver. & Fair Aver.	Raw score	Fair score	Difference in total score
23	29	12	2.42	2.78	-0.88	0.58	1.52	1.55	-0.36	9.7	11.1	-1.4
25	30	12	2.5	2.85	-0.55	0.57	0.93	0.92	-0.35	10.0	11.4	-1.4
19	28	8	3.5	3.83	3.76	0.68	0.74	0.73	-0.33	14.0	15.3	-1.3
20	28	8	3.5	3.83	3.76	0.68	0.65	0.64	-0.33	14.0	15.3	-1.3
76	28	8	3.5	3.79	3.6	0.7	2.08	2.02	-0.29	14.0	15.2	-1.2
75	29	8	3.63	3.91	4.09	0.7	0.82	0.83	-0.28	14.5	15.6	-1.1
71	27	8	3.38	3.66	3.1	0.72	1.81	1.61	-0.28	13.5	14.6	-1.1
74	27	8	3.38	3.66	3.1	0.72	0.49	0.44	-0.28	13.5	14.6	-1.1
73	26	8	3.25	3.52	2.56	0.74	0.73	0.62	-0.27	13.0	14.1	-1.1
21	34	12	2.83	3.1	0.83	0.63	0.58	0.55	-0.27	11.3	12.4	-1.1
22	24	8	3	3.25	1.53	0.83	0.03	0.03	-0.25	12.0	13.0	-1.0
24	24	8	3	3.25	1.53	0.83	0.03	0.03	-0.25	12.0	13.0	-1.0
67	35	12	2.92	2.68	-1.28	0.65	1.12	1.15	0.24	11.7	10.7	1.0
192	30	8	3.75	3.51	2.54	0.71	0.95	0.95	0.24	15.0	14.0	1.0
189	28	8	3.5	3.26	1.58	0.69	0.88	0.88	0.24	14.0	13.0	1.0
190	29	8	3.63	3.38	2.05	0.69	0.64	0.62	0.25	14.5	13.5	1.0
191	29	8	3.63	3.38	2.05	0.69	1.33	1.34	0.25	14.5	13.5	1.0
143	28	8	3.5	3.25	1.53	0.68	0.9	0.9	0.25	14.0	13.0	1.0
146	28	8	3.5	3.25	1.53	0.68	0.9	0.9	0.25	14.0	13.0	1.0
58	33	12	2.75	2.48	-2.03	0.61	1.5	1.62	0.27	11.0	9.9	1.1
70	30	12	2.5	2.23	-3.11	0.58	0.78	0.76	0.27	10.0	8.9	1.1
68	33	12	2.75	2.47	-2.07	0.61	1.5	1.6	0.28	11.0	9.9	1.1
60	43	12	3.58	3.29	1.72	0.56	1.65	1.72	0.29	14.3	13.2	1.2

RATER ANALYSIS: PRACTICAL USE

- **MFRM can evaluate all facets**
  - Rater performance
  - Rating scale performance
  - Student performance
- **MFRM can help identify**
  - Consistent raters and inconsistent raters
  - Appropriate, harsh or lenient markers
  - Raters with instances of unexpected severe/lenient ratings
  - Criteria that fit or do not fit the model
  - Criteria are harder to mark accurately

RATER ANALYSES: FEEDBACK AT MUELC

- Inform all raters of their rating performance
- Specify next steps
  - ✓ **Consistent and appropriate** raters:
    - ❑ continue to refer to self-access sample bank
    - ❑ do required online rater training tasks before next marking period
  - ✓ **Inconsistent, lenient, harsh** raters:
    - ❑ continue to refer to self-access sample and benchmark bank
    - ❑ do required online rater training tasks before next marking period
    - ❑ **attend face-to-face rater training**

RATER ANALYSES: RATER TRAINING AT MUELC

Message: to **support teachers** in their growth as raters and teachers

Set up of targeted rater training

- Target criteria that are harder to mark consistently and appropriately
- Go through online samples, awarded scores and benchmark comments
- Raters mark a sample on the spot and discuss scores
- Raters reflect on previous marking behaviours and align their scores via the use of the rating scale.
- Give feedback to raters during the training process

THANK YOU ALL FOR YOUR ATTENTION!

REFERENCES

- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328. Retrieved from http://www.rasch.org/rmt/rmt74m.htm.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm.
- Linacre, J. M. (2015). *Facets computer program for many-facet Rasch measurement*, Version 3.71.4. Beaverton, Oregon: Winsteps.com.
- Myford, C. M., & Wolfe, E. W. (2004a). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460 – 517). Maple Grove, MN: JAM Press.
- Myford, C. M., & Wolfe, E. W. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 518 – 574). Maple Grove, MN: JAM Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8(3)*, 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm.