UNIVERSITY of CAMBRIDGE
ESOL Examinations
English for Speakers of Other Languages

COUNCIL OF EUROPE CONSEIL DE L'EUROPE

Seminar to calibrate examples of spoken performance
CIEP Sèvres, 02-04.12.2004

Report on analysis of rating data

Dr Neil Jones
Senior Research and Validation Coordinator
Cambridge ESOL

March 1, 2005

Introduction

This report is based on the electronically-captured ratings of 38 raters, rating 25 video extracts of spoken performance.
The six rating criteria used were:

Etendue	Range
Correction	Accuracy
Aisance	Fluency
Interaction	Interaction
Cohérence	Coherence
Note_Global	Global rating

There were two rating modes: independent, that is, after discussion in a limited group, and after plenary discussion. Not all extracts were rated on the full set of criteria in both modes; in particular, the after discussion ratings omitted interaction and coherence.
Apart from this the matrix of ratings is fairly complete, that is, nearly all raters rated all extracts.

The following questions are addressed in this report:

1. What is the best estimate of the CEFR level of each extract?
2. How well do raters agree in their ratings?
3. What is the effect of plenary discussion on the extent of agreement?
4. How do raters understand and use the rating criteria?
5. Does agreement improve over time?
6. Do rater groups perform differently?

1.1 Data formatting and analysis

Data were supplied in a number of EXCEL sheets produced by the electronic voting system. These were imported into an ACCESS database together with other tables:

1. The table of raters. Raters were grouped into four categories:

Institutions d'évaluation de langue française	Assessment bodies for French
Écoles de langue française (en France)	French language schools (in France)
Projets européens et experts de langue française	European projects and French language experts
Projets européens: autres participants	European projects, other participants

2. The table of subjects, including the day on which each subject was rated (Thursday, Friday, Saturday).
3. The table of voting categories, each identified as a combination of a rating criterion, as identified as above, and a rating mode, i.e. independent or after discussion.

Queries were run in ACCESS to facilitate construction of a data file in the format required by FACETS for multi-faceted Rasch analysis. An example extract of one such file is shown in 4.2 below.
A number of summary analyses were done as queries in ACCESS, in conjunction with EXCEL to produce graphs.
A generalizability study was also run using GENOVA.

1.1.1 Multifaceted Rasch measurement

Many-facet Rasch measurement extends the simple Rasch model by allowing the difficulty of achieving a certain score on a test task to be decomposed into separate facets, each of which can be separately

estimated. Thus possible facets in the present data include the difficulty of each rating criterion, the severity of raters, and any effect connected with the group to which they belonged. The software used is FACETS (Linacre ---).

1.1.2 Generalizability theory

In classical test analysis the reliability of a test is defined as the ratio of true-score variance to total variance. Total variance is always the greater, because it includes error variance. There may be many sources of measurement error, but these are not differentiated in classical test analysis. Generalizability theory enables separate sources of error to be identified, and thus enables particular conditions of test administration to be modeled, and the reliability under those conditions estimated. Thus in the present study it is possible to model the reliability of ratings made by different numbers of raters, using different numbers of rating criteria. The software used is GENOVA (Brennan and Crick 19--).

2 Findings and discussion

2.1 What is the best estimate of the CEFR level of each extract?

A key aim of the seminar was to agree definitive levels for the extracts. Several methods were compared, and a final decision made by finding a best match across these methods. The methods included:

1. Ways of summarizing the raw data, i.e. the mean rating across rating criteria and raters, rounded to the nearest level, and the modal, i.e. the most frequent rating for each extract;
2. Interpreting results from different FACETS analyses, e.g. of independent ratings, of ratings made after discussion, anchored to CEFR scale cutoffs or unanchored.

In fact the question of how to derive the level of an extract from a FACETS analysis turns out to be quite a difficult one. The issues seem to require some detailed discussion, which I have relegated to an appendix (4.2.1 below). They are reviewed more briefly in the Conclusions (3 below).

2.2 How well do raters agree in their ratings and what is the effect of plenary discussion on the extent of agreement?

The following indices of agreement are provided:

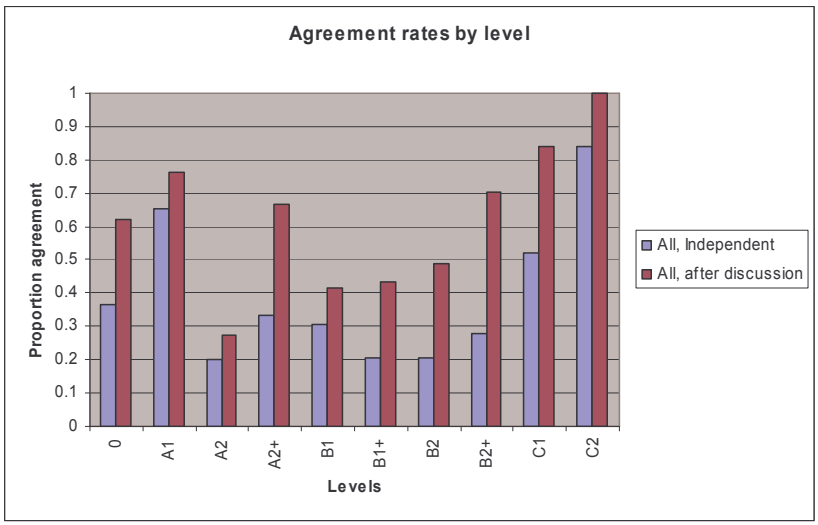
1. Raw agreement indices
2. Graphical representations of agreement
3. Inter-rater correlations.
4. FACETS statistics (severity, fit, reliability of subject ability estimate)
5. A generalizability study

The kappa coefficient (Cohen 1960) has been suggested as an index of agreement. However, it appears that its use is contentious (e.g. Thompson and Walter 1988, further discussion in Uebersax 2002b), and I have not included it here.

2.2.1 Raw agreement indices

Raw agreement indices were calculated using the generalized case formulae given by Uebersax (2002a). These allow observed agreement to be summarized as a proportion of the theoretically possible agreement. Figure 1 illustrates.

Figure 1 Agreement rates by level before and after plenary discussion



Agreement would be 100% if all raters used the same rating category for each subject. Agreement is higher to the extent that raters use a smaller number of rating categories.

Figure 1 shows that agreement is highest for the extreme levels A1 and C2, and tends to be lowest towards the middle of the scale. It shows that agreement is generally much better after plenary discussion, reaching 100% in the case of the two C2 subjects.

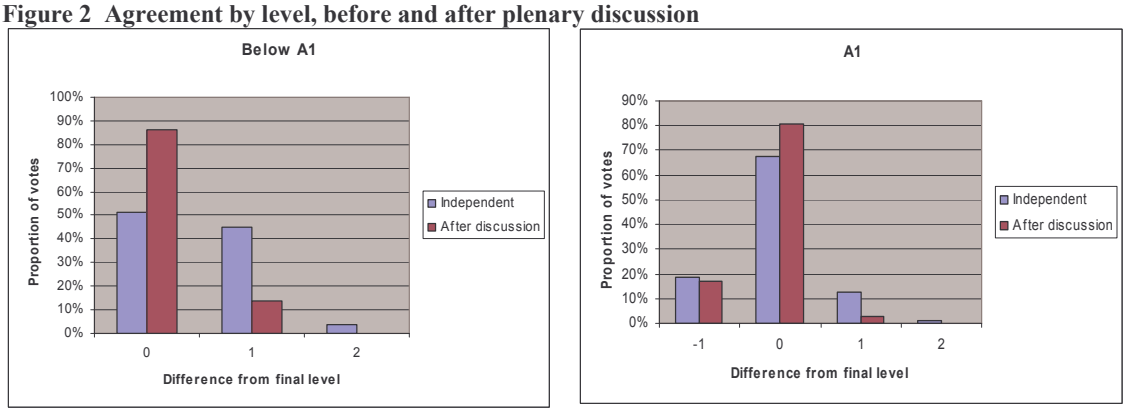
Figure 1 shows the agreement rate estimated across all raters and ratings. For some FACETS analyses six raters were removed because of misfit. Agreement rates were also estimated for this reduced rater group, but differences were small:

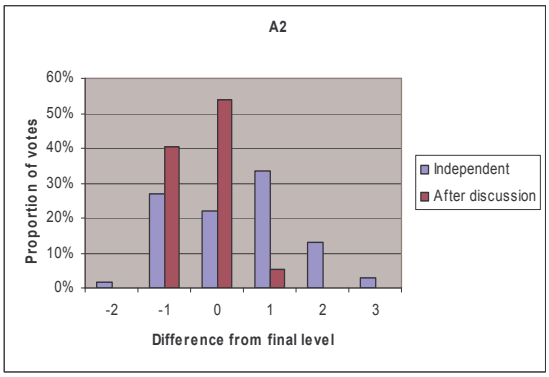
Overall agreement	All	6 raters removed
Independent	39%	40%
After discussion	64%	66%

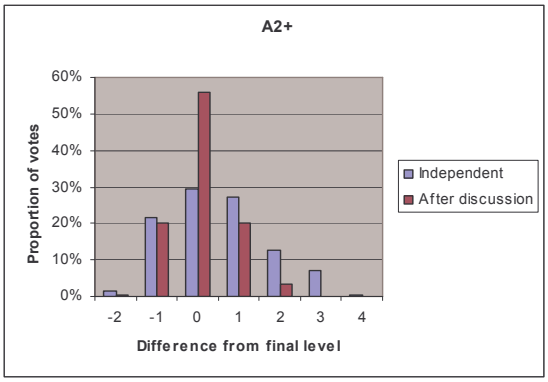
2.2.2 Graphical representations of agreement by level

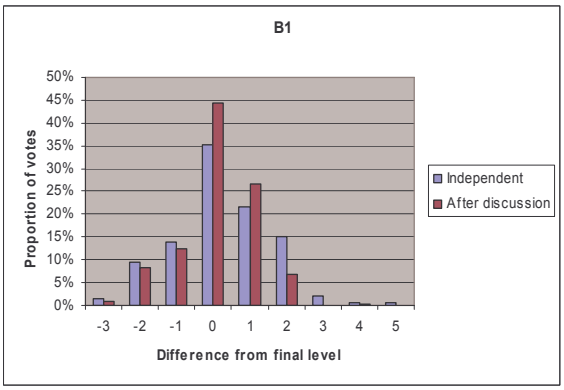
The following graphs give a representation of agreement by level, before and after plenary discussion. They show disagreements as deviations from the agreed rating (using the modal rating, i.e. this may not correspond exactly to the final assignment to level).

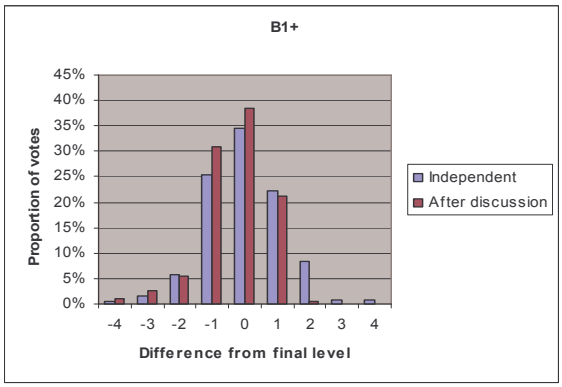
Figure 2 Agreement by level, before and after plenary discussion

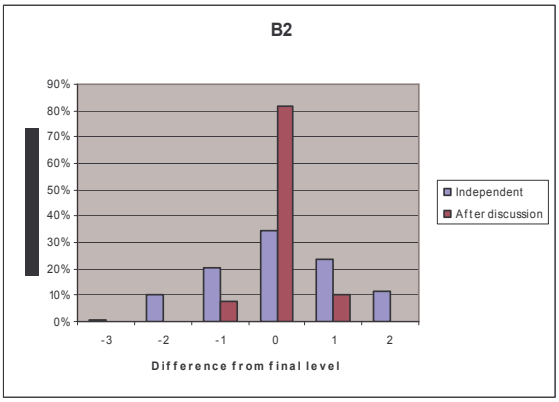


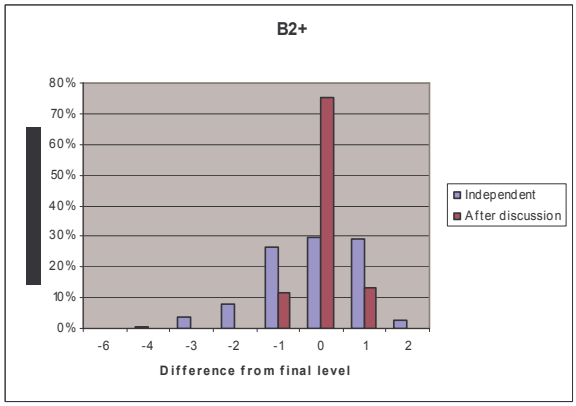

A2
Proportion of votes
60%
50%
40%
30%
20%
10%
0%
-2
-1
0
1
2
3
Difference from final level
Independent
After discussion

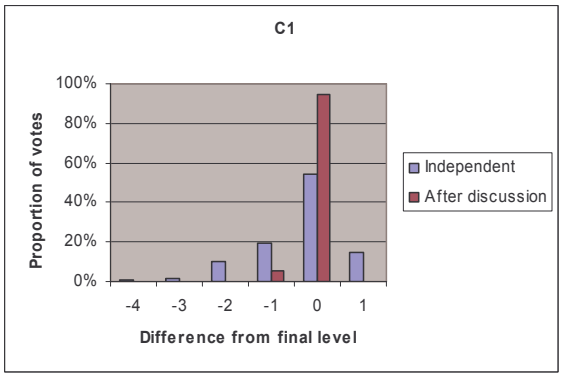

A2+
Proportion of votes
60%
50%
40%
30%
20%
10%
0%
-2
-1
0
1
2
3
4
Difference from final level
Independent
After discussion

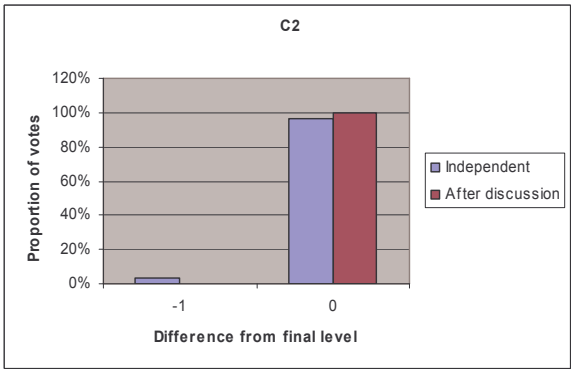

B1
Proportion of votes
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
-3
-2
-1
0
1
2
3
4
5
Difference from final level
Independent
After discussion


B1+
Proportion of votes
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
-4
-3
-2
-1
0
1
2
3
4
Difference from final level
Independent
After discussion


B2
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
-3
-2
-1
0
1
2
Difference from final level
Independent
After discussion


B2+
80%
70%
60%
50%
40%
30%
20%
10%
0%
-6
-4
-3
-2
-1
0
1
2
Difference from final level
Independent
After discussion


C1
Proportion of votes
100%
80%
60%
40%
20%
0%
-4
-3
-2
-1
0
1
Difference from final level
Independent
After discussion


C2
Proportion of votes
120%
100%
80%
60%
40%
20%
0%
-1
0
Difference from final level
Independent
After discussion

2.2.3 Correlation as an index of agreement

Mean inter-rater correlations were calculated, using complete data only, i.e. omitting cases where some raters did not provide a rating.

Table 1 Mean Inter-rater correlation

	r
Independent	0.886
After discussion	0.967

2.2.4 FACETS statistics

Variation in severity

As indicated by the FACETS output, (e.g. Figure 7 below), the raters vary in severity within a narrow range, relative to the ability range of the subjects. (the SD ratio is 8.5 : 1 for an analysis including all raters, and over 12 : 1 for an analysis omitting 3 most lenient raters.)

Figure 3 Distribution of rater severity (all raters)

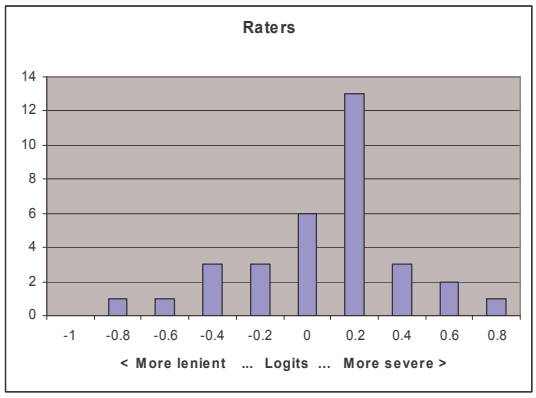


Figure 3 shows that the distribution of rater measures is slightly negatively skewed, that is, there is a small tail of rather lenient raters. However, removing the three most lenient raters has no effect on the grading of subjects in an anchored analysis.

Variation in fit

Fit statistics for raters give an indication of whether individuals rate generally consistently with others. Rater fit statistics for this dataset are generally satisfactory: six slightly misfitting raters were removed from calibration runs in order to minimize misfit, but the effect is fairly minimal.

Reliability of subject ability estimates

These are high: r = 1.00 (FACETS reports to 2 dp) for the independent ratings.

It is worth noting that the FACETS analyses of the ratings made *after* plenary discussion are rather unsatisfactory for the purposes of scale construction. The Rasch model depends on the local independence of ratings, a condition clearly violated by the nature of the plenary discussion, the purpose of which is to achieve something approaching unanimity. In consequence the analysis of ratings after discussion shows a great deal of overfit and one scale category (9, or C2) even becomes unmeasurable, because there are no instances of disagreement in the use of that category. It is crucial that if part of the purpose of a rating exercise is to construct a measurement scale then there must be an opportunity for rating both before and after plenary discussion.

2.2.5 Generalizability study

A generalizability study was done using the program GENOVA. Data were the independent ratings for the 17 subjects and 35 raters who constituted a balanced design (no missing data). G-theory allows the reliability of rating to be estimated where fewer raters are used and perhaps fewer rating criteria. The generalizability coefficients found for different combinations are shown in Table 2 below.

Table 2 Generalizability coefficients for different numbers of raters and criteria

	No. of rating criteria	
No. of raters	6	1
35	0.997	0.993
2	0.962	0.949
1	0.928	0.905

The generalizability for all raters is very high and compares with the reliability of 1.00 reported by FACETS (reported to 2 dp) for the reliability of the subject estimates.
Reducing the number of rating criteria to one has very little effect on this, because the rating criteria are highly correlated (see 2.3 below).

When the number of raters is reduced to two or even one the reliability remains high. This must reflect among other things the wide range of ability (A1 to C2) encompassed.

As Linacre (1993) observes, the Rasch model enables generalizability to be estimated directly, based on relationships between the SD of ability in the population, and the number and kind of ratings (dichotomies or rating scales with more or fewer categories). The nomograph provided by Linacre suggests similar values to those shown above.

2.3 How do raters understand and use the rating criteria?

Are the rating criteria understood and interpreted in distinct ways?
This question is addressed by:

1. Correlation of ratings on each criterion;
2. Raw agreement indices by rating criterion
3. Fit statistics from FACETS analysis;
4. Difficulty statistics from FACETS analysis

2.3.1 Correlation of ratings on each criterion

Correlations between ratings on different performance criteria will be lower if distinct aspects of performance are picked up under each criterion (assuming that the subjects also have distinct profiles of skill). Correlations can thus be used to test just how distinct criteria are in the way they are used. Correlations must be disattenuated (see 4.2.1 below for what this means). If a disattenuated correlation is still less than 1 then this can be taken as evidence that the correlated traits are measuring distinct skills.

The abilities of all subjects were estimated in FACETS runs for each rating criterion separately (independent ratings only, misfitting raters removed). The reliability of each rating was taken as that reported by FACETS: .99 for every criterion.
These independent estimates were correlated. In fact the correlations are very high (Table 3) and when disattenuated reach 1.

Table 3 Correlation of independently-estimated abilities on each rating criterion

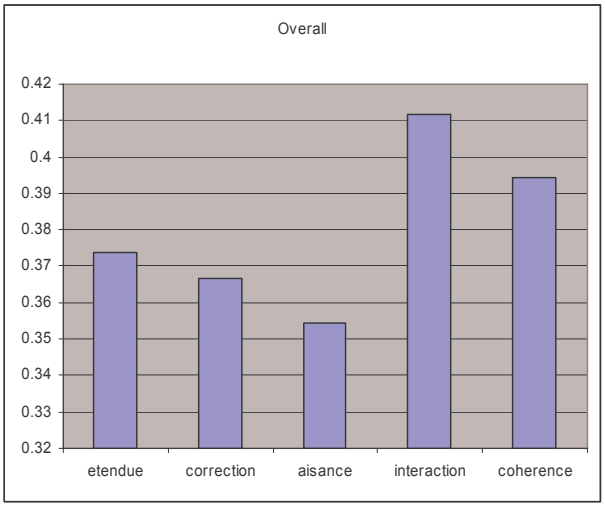
	Crit1	*Crit2*	*Crit3*	*Crit4*	*Crit5*
Crit1	1				
Crit2	0.998	1			
Crit3	0.997	0.995	1		
Crit4	0.994	0.990	0.992	1	
Crit5	0.998	0.997	0.998	0.992	1

This correlation based on estimated abilities was repeated using raw ratings, producing very similar results.
On this evidence either the rating criteria are not distinct, or the raters were not able to identify that they were distinct (=halo effect) or the subjects do not display distinct profiles of skill.

2.3.2 Raw agreement indices by rating criterion

Figure 4 shows the proportion of agreement by rating criterion. Agreement varies from about 35% for *aisance* (fluency) to 41% for interaction. This suggests a small difference in agreement across criteria.

Figure 4 Proportion of agreement by rating criterion



2.3.3 Fit statistics for rating criteria

Table 4 shows the fit statistics for the rating criteria. Misfit (positive values) suggests relatively less consistent use of a category, while overfit (negative values) suggests greater consistency. Here *correction* (accuracy) seems to attract slightly less agreement. Unsurprisingly the global mark overfits strongly, as if effectively averages the other criteria.

Table 4 Standardised Infit of rating criteria from FACETS analysis

	Infit ZStd
Etendue	0
Correction	2
Aisance	0
Interaction	0
Cohérence	-1
N_Global	-4

2.3.4 Difficulty of rating criteria

The rating criteria differ slightly in difficulty, but the difference is significant (r = .96).

Figure 5 Relative difficulty of rating criteria (FACETS analysis)

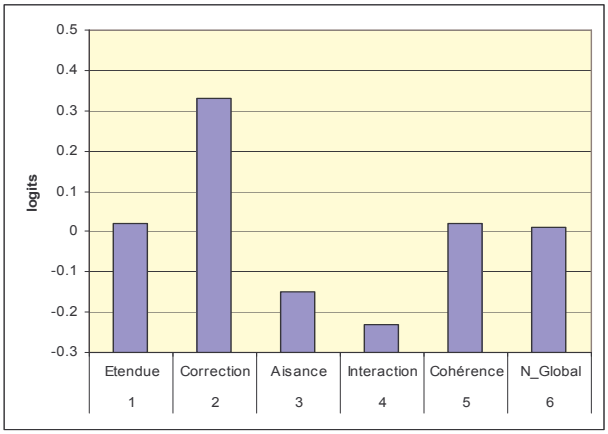


Figure 5 shows that *Correction* (accuracy) is most severely rated, and Interaction most leniently.

2.4 Does agreement improve over time?

It is difficult to make a comparison over the three days of the seminar to the extent that the focus of activity was somewhat different each day, and because the number of subjects rated, and their ability range, varied. However, one would expect that other things being equal agreement rates would improve over time if the purpose of achieving a shared understanding of levels were achieved.

2.4.1 Raw agreement indices

Table 5 shows the overall agreement on each day.

Table 5 Proportion of agreement on three days

	Overall
Day 1	25%
Day 2	44%
Day 3	33%

Table 6 shows the fit statistics for the Days facet of a FACETS analysis. Misfit (positive values) suggests relatively less consistent rating behaviour on a particular day. Consistency appears to improve over the three days.

Table 6 Standardised infit for Days (FACETS analysis)

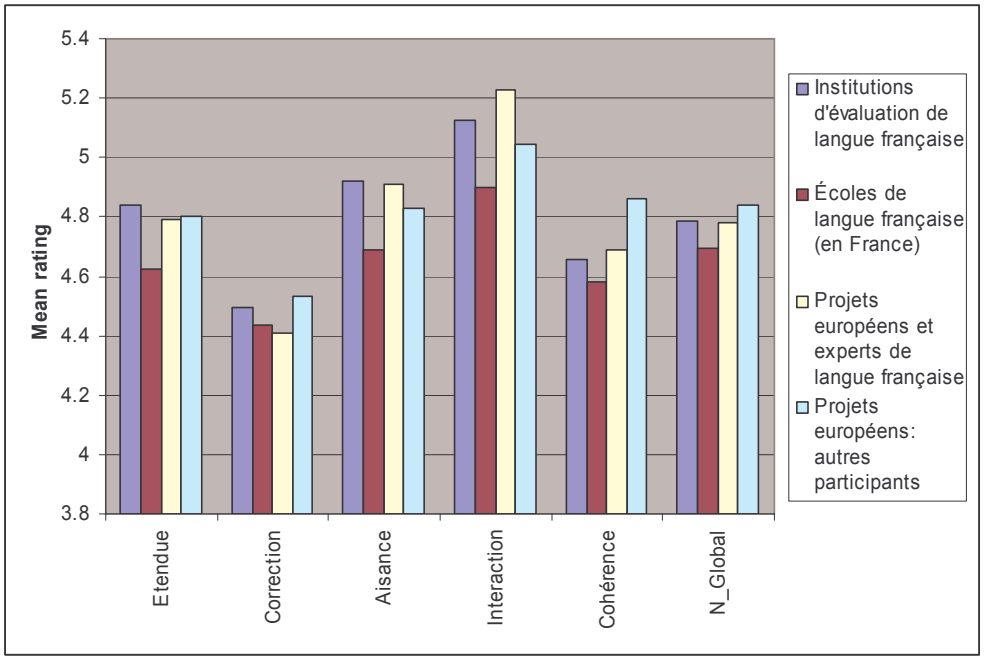
	Infit ZStd
Day 1	7
Day 2	4
Day 3	-2

2.5 Do rater groups perform differently?

2.5.1 Mean rating by criterion and by rater group

Overall the four groups identified are of very similar severity. Figure 6 suggests some possible minor differences. The *European projects and French language experts* group appears to use a slightly wider range, being the most severe on *Correction* (accuracy) but the most generous on Interaction. This contrasts with the *French language schools (in France)* group, which rates more evenly across the criteria.

Figure 6 Mean rating by criterion and by rater group (independent ratings)



Raw indices of agreement by rater group and by rating criterion were computed. They do not show any clear differences in the way different groups use particular criteria.

3 Conclusions

The Sevres seminar provides an excellent model for how similar events should be run in future. Important features included:

1. The participation of a large number of well-qualified language professionals, including those with authority to interpret the video extracts in relation to the CEFR,
2. A technical infrastructure (electronic voting) which enabled the error-free capture of a large amount of data;
3. A planned sequence of stages, beginning with familiarisation with the rating criteria.

Inevitably perhaps the approach was revised during the event. My interpretation of this process is that:

1. the early ratings were felt to be excessively individualistic, failing to create the feeling of a group endeavour to create a shared understanding of the levels;
2. the remedy for this was to focus more on the text of the CEFR descriptors, that is, to present a more authoritative view of how particular extracts should be interpreted in relation to them;
3. in consequence the standardisation aspect, focussing on the interpretation of scale descriptors, took prominence over a more exploratory, construct-focussed approach based on the direct comparison of performances.

It may be that it is difficult or impossible to pursue both goals within the context of a single event. In this case the emphasis placed on standardisation was certainly the correct one, given the requirement to produce a set of accurately-rated exemplars for the DVD planned as support material for the CEFR Manual.

None the less, the direct comparison of performances remains an important goal, because equating levels across languages can probably only be done satisfactorily by suitably qualified judges making such direct comparisons. Comparisons mediated by the text of the CEFR are indirect and subject to a problem evident in the seminar, that the wording of the illustrative scales is to an extent relative rather than absolute.

It was very important that two sets of ratings were captured, both before and after discussion. The discussion in small groups was particularly useful because this seemed to favour useful reflexion on features of the performance and on why one was inclined to a particular rating. The plenary discussion that followed the reports from groups and precede the final vote tended to have a more prescriptive tone. The difference is evident in the data: the data which are amenable to analysis using FACETS are the independent ratings, because they contain enough variability to enable scale construction. The post-plenary data provide the best view of the final consensus as to the level of each extract, but are quite unsatisfactory from the point of view of Rasch analysis, because of the evident lack of local independence in the ratings – they strongly violate the assumptions of the Rasch model. It is crucial that if part of the purpose of a rating exercise is to construct a credible measurement scale then there must be an opportunity for rating both before and after plenary discussion. It is of course most satisfactory when the final consensus decision after the discussion matches the level of each extract as defined by the FACETS analysis of the independent ratings.

One issue of possible significance is the use of the different rating criteria. In this dataset it is clear that use of the criteria by raters does *not* identify different, distinct profiles of ability – they correlate too highly for that. At the same time they are rated more or less severely: raters are tougher on accuracy and more generous with respect to interaction. Are the raters wrong here, that is, are they misinterpreting the CEFR descriptors? Or are they right, in which case we must ask, why are the CEFR descriptors worded as they are? Shouldn't B1 typically be evidenced by B1 performance across all the rating criteria?

The analysis of rater agreement given above shows that agreement is easier in the case of the extremes of the ability scale and harder around the middle. Is this just a fact of life – a simple consequence of the particular distribution of abilities that we typically have to deal with in teaching or assessment? Or is B1 intrinsically a less clearly definable level than A1 or C1 - one where learners may exhibit very different profiles, and features of a range of levels? There might be practical implications for approaches to rater training.

Finally, how best to use the ratings collected during the seminar to assign authoritative levels to the extracts? As discussed in Part 2.1 above and in more detail in 4.2.1 below, FACETS analysis is part of the answer – less intuitive perhaps than simply taking mean or most frequent ratings, but superior to the extent that it should be able to factor out effects caused by differential rater severity or use of rating criteria. However, FACETS offers three alternative approaches to defining levels: *modal*, *median* and *mean*. The modal scale is attractive for two reasons:

1. its meaning – the range where a given rating is the most probable, or frequent – makes intuitive sense;
2. it is what is specified when scale steps are anchored, i.e. it reflects an intention that subjects falling within specified ranges should be assigned to certain levels.

However, there are examples in the present data where the modal scale produces counter-intuitive level assignments. In the present analysis the median (or mean) appears more dependable. Given the great usefulness of FACETS for analysing the kind of data produced at rating events like the Sevres seminar it is important that a common principle for interpretation be agreed.

4 Appendices

4.1 Final level of subjects

Table 7 shows the final decision as to the level of the subjects rated during the conference. The final level is a subjective decision based on comparing a number of different indices – summarized raw scores, as well as levels from FACETS analyses using mean, median and modal scale steps. The two FACETS columns in Table 7 are based on median scale steps.

Table 7 Final level of subjects

		Independent Judgements (FACETS)	Judgements after Discussion (FACETS)	**Definitive CEF Level**
33	Josue	C2	C2	**C2**
32	Rachel	C2	C2	**C2**
31	Aleksandar DALF	C1	C1	**C1**
15	Ambriogio	C1	C1	**C1**
25	Aleksandar	C1	C1	**C1**
8	Xi	B2+	B2+	**B2+**
26	*Luis*	*B2+*	*B2+*	***(B2+)***
16	Silvia	B2+	B2+	**B2+**
7	Nataliya	B2	B2	**B2**
37	Gu Jung	B1+	B1+	**B1+**
4	Sophie	B1+	B1+	**B1+**
3	Valérie	B1+	B1	**B1**
13	Evelyne	B1	B1	**B1**
1	Margarida	B1	B1	**B1**
2	Mariana	B1	B1	**B1**
14	Andrea	B1	A2+	**(A2+) / B1**
5	Debora	B1	A2+	**A2+**
61	Katell	A2+	A2+	**A2+**
38	Aamer	A2+	A2+	**A2+**
6	Iryna	A2+	A2+	**A2+**
62	Sun Ying	A2+	A2	**A2**
50	Viggo	A1 (high)	A1 (high)	**A1**
42	Suzanne	0	A1	**A1**
41	Sally	0	A1	**A1**
49	Jessica	0	0	**0**

4.2 Analysis and methodology

4.2.1 Using FACETS to derive level estimates

It is worth discussing the FACETS analysis in some detail to make it clear how it establishes the level of a subject relative to the scale cutoffs. The following discussion and illustrations refer to the FACETS analysis of the *independent* ratings.

Figure 7 shows part of the FACETS output: Table 6, which arranges the different facets of the analysis vertically against a logit measurement scale (the first column on the left). The subjects are arranged from the highest-ability (Josue and Rachel) to the lowest (Jessica). The raters are arranged by severity, from the most severe to the most lenient (the range is relatively narrow). The next column "Scale" shows the

rating criteria. Criterion 2 (Correction) is very slightly harder (rated more severely) and Criterion 4 (Interaction) is slightly easier.
The column on the far right shows how the rating scale has been used. The raters voted using CEFR categories A1 – C2, including plus levels, but these have been converted to integers for the analysis. Thus 1 = A1, 2 = A2, 3 = A2+ and so on.
When the analysis estimates the ability of each subject it compensates for the effect on ratings of rater severity or the difficulty of a given criterion, by adjusting their average rating to give a "fair average". The vertical position of a subject corresponds to their fair average score, which is indicated by the right-hand column. The bars between the numbers correspond to a half-mark, i.e. 1.5, 2.5, 3.5 etc.
Thus for example Nataliya has a fair average score of 6.01, which places her directly opposite the "6" on the scale.

The next thing is to understand how these fair average scores relate to levels. Does a fair average of 5.7 put you at Level 5 or Level 6, for example? FACETS in fact has three different ways of interpreting a fair average in terms of a level, which are referred to as the *modal*, the *median* and the *mean*. The scale in the FACETS vertical summary shows the mean, but this can vary considerably from the median and modal.

Figure 7 shows an unanchored analysis. The modal, mean and median values found in this analysis are shown in the graph at the foot of the figure. As suggested by the dotted lines above each group of columns, the relationship of modal and mean differs at each level. Thus for example the modal difference between levels 3 and 4 is much smaller than the mean difference.

Table 8 shows the effects on the estimated level of subjects using the three different values from this unanchored analysis. The subjects are shown in descending order of ability. Using the modal values no subjects fall within level A2+, because the interval between modal A2+ and B1 is very narrow. On the other hand no less than nine subjects fall at B1, because the interval between modal B1 and B1+ is much wider. A different picture is given if the median or mean values are used.

Figure 7 Example FACETS vertical summary, unanchored analysis

```
Sevres Jan 25/05 Minus 6 raters, NO anchor to cef levels  01-29-2005 14:21:17
Table 6.0  All Facet Vertical "Rulers".
-------------------------------------------------------------------------------------------
|Measr|+sujets                                          |-raters    |-Scale   |-Indep  |S.1  |
-------------------------------------------------------------------------------------------
+   8 + Josue                                          +           +         +        +(9)  +
|     | Rachel                                         |           |         |        |     |
|     |                                                |           |         |        |     |
+   7 +                                                +           +         +        +     +
|     |                                                |           |         |        |     |
|     |                                                |           |         |        |     |
+   6 +                                                +           +         +        +     +
|     |                                                |           |         |        |     |
|     |                                                |           |         |        |     |
+   5 +                                                +           +         +        + --- +
|     |                                                |           |         |        |     |
|     |                                                |           |         |        |     |
+   4 + Aleksandar DALF                                +           +         +        + 8   +
|     |                                                |           |         |        |     |
|     | Ambriogio                                      |           |         |        |     |
+   3 + Aleksandar                                     +           +         +        + --- +
|     |                                                |           |         |        |     |
|     | Luis             Silvia     Xi                 |           |         |        | 7   |
+   2 +                                                +           +         +        + --- +
|     | Nataliya                                       |           |         |        | 6   |
|     |                                                |           |         |        | --- |
+   1 + Gu Jung          Sophie                        +           +         +        +     +
|     | Valérie                                        | *         |         |        | 5   |
|     | Evelyne          Margarida                     | ***       | 2       |        | --- |
*   0 * Debora           Mariana                       * ********. * 1 3 5 6 * *      * 4   *
|     | Andrea                                         | **.       | 4       |        | --- |
|     | Aamer            Iryna      Katell  Sun Ying   | *.        |         |        | 3   |
+  -1 +                                                +           +         +        +     +
|     |                                                |           |         |        | --- |
|     |                                                |           |         |        |     |
+  -2 +                                                +           +         +        + 2   +
|     |                                                |           |         |        |     |
|     |                                                |           |         |        |     |
+  -3 +                                                +           +         +        +     +
|     | Viggo                                          |           |         |        | --- |
|     |                                                |           |         |        |     |
+  -4 +                                                +           +         +        +     +
|     |                                                |           |         |        |     |
|     |                                                |           |         |        |     |
+  -5 +                                                +           +         +        + 1   +
|     |                                                |           |         |        |     |
|     |                                                |           |         |        |     |
+  -6 + Sally           Suzanne                       +           +         +        +     +
|     |                                                |           |         |        |     |
|     |                                                |           |         |        |     |
+  -7 + Jessica                                       +           +         +        +(0)  +
-------------------------------------------------------------------------------------------
|Measr|+sujets                                          | * = 2     |-Scale   | * = 1  |S.1  |
```

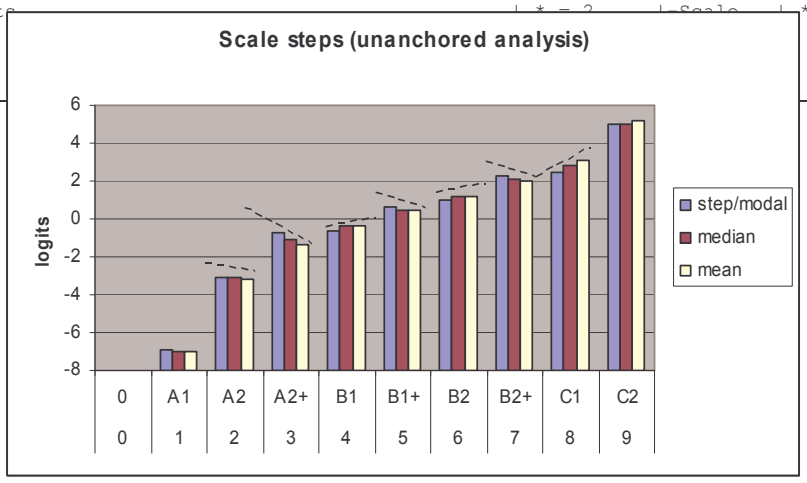


Table 8 Unanchored analysis: Levels derived from modal, median and mean cutoffs

	UNANCHORED	fair average	ability	step/modal	median	mean
33	Josue		10.96	C2	C2	C2
32	Rachel	8.9	7.64	C2	C2	C2
31	Aleksandar DALF	8.1	4	C1	C1	C1
15	Ambriogio	7.6	3.2	C1	C1	C1
25	Aleksandar	7.5	3.05	C1	C1	C1
8	Xi	7	2.47	C1	B2+	B2+
26	Luis	6.9	2.4	B2+	B2+	B2+
16	Silvia	6.7	2.2	B2	B2+	B2+
7	Nataliya	6.1	1.66	B2	B2	B2
37	Gu Jung	5.1	0.9	B1+	B1+	B1+
4	Sophie	5.1	0.88	B1+	B1+	B1+
3	Valérie	4.6	0.51	B1	B1+	B1+
13	Evelyne	4.5	0.4	B1	B1	B1
1	Margarida	4.2	0.21	B1	B1	B1
5	Debora	3.9	-0.03	B1	B1	B1
2	Mariana	3.8	-0.15	B1	B1	B1
14	Andrea	3.7	-0.2	B1	B1	B1
61	Katell	3.3	-0.51	B1	A2+	A2+
38	Aamer	3.3	-0.52	B1	A2+	A2+
62	Sun Ying	3.2	-0.6	B1	A2+	A2+
6	Iryna	3	-0.83	A2	A2+	A2+
50	Viggo	1.4	-3.47	A1	A1	A1
42	Suzanne	0.8	-5.93	A1	A1	A1
41	Sally	0.8	-5.97	A1	A1	A1
49	Jessica	0.5	-6.92	A1	A1	A1

Let us compare Figure 7 with a second analysis in which the scale steps are anchored.
The modal levels identify the ability ranges where a given rating is found to be the most probable, i.e. should be most frequently observed. The modal levels are identical to the step difficulty calibrations, which in an anchored analysis are constrained to take on specified logit values. In this study some analyses were anchored to values previously specified for the CEFR levels (North 2000: 274). This has the advantage that the abilities estimated for the extracts are on the same scale as the illustrative scales published in the CEFR.
Figure 8 shows such an anchored analysis. The scale steps are fixed at nearly equal intervals, which is reflected in the appearance of the vertical scale and the consequent re-alignment of the subjects. The graph at the foot of the figure shows that the modal and mean cutoffs are very similar, with the notable exception of the lowest and highest cutoffs.
Table 9 shows the effects on the estimated level of subjects using the three different values. They give very similar results, with the exception of the bottom of the scale, where the modal values put Sally and Suzanne at level 0 rather than A1.

Figure 8 Example FACETS vertical summary, analysis with scale steps anchored to CEFR cutoffs

```
Sevres Jan 25/05 Minus 6 raters, anchor to cef levels  01-30-2005 10:35:19
Table 6.0  All Facet Vertical "Rulers".
-------------------------------------------------------------------------------
|Measr|+sujets                          |-raters |-Scale |-Independent|S.1   |
-------------------------------------------------------------------------------
+   7 + Josue                           +        +       +            +(9)   +
|     |                                 |        |       |            |      |
|     | Rachel                          |        |       |            |      |
|     |                                 |        |       |            |      |
+   6 +                                 +        +       +            +      +
|     |                                 |        |       |            |      |
|     |                                 |        |       |            |      |
|     |                                 |        |       |            |      |
+   5 +                                 +        +       +            +      +
|     |                                 |        |       |            |      |
|     |                                 |        |       |            | ---  |
|     |                                 |        |       |            |      |
+   4 +                                 +        +       +            +      +
|     |                                 |        |       |            |      |
|     | Aleksandar DALF                 |        |       |            | 8    |
|     |                                 |        |       |            |      |
+   3 + Ambriogio                       +        +       +            +      +
|     | Aleksandar                      |        |       |            | ---  |
|     |                                 |        |       |            |      |
|     | Luis             Xi             |        |       |            | 7    |
+   2 + Silvia                          +        +       +            +      +
|     |                                 |        |       |            | ---  |
|     |                                 |        |       |            |      |
|     | Nataliya                        |        |       |            | 6    |
+   1 +                                 +        +       +            +      +
|     |                                 |        |       |            | ---  |
|     |                                 | *.     |       |            |      |
|     | Gu Jung          Sophie         | ****.  | 2     |            | 5    |
*   0 *                                 * ****** * 1 5 6 * *          *      *
|     | Evelyne          Valérie        | **     | 3 4   |            | ---  |
|     | Margarida                       | *.     |       |            |      |
|     | Debora                          | *      |       |            | 4    |
+  -1 + Andrea           Mariana        +        +       +            +      +
|     | Katell                          |        |       |            | ---  |
|     | Aamer            Sun Ying       |        |       |            |      |
|     | Iryna                           |        |       |            | 3    |
+  -2 +                                 +        +       +            +      +
|     |                                 |        |       |            | ---  |
|     |                                 |        |       |            |      |
|     |                                 |        |       |            | 2    |
+  -3 +                                 +        +       +            +      +
|     | Viggo                           |        |       |            | ---  |
|     |                                 |        |       |            |      |
|     |                                 |        |       |            | 1    |
+  -4 +                                 +        +       +            +      +
|     | Sally            Suzanne        |        |       |            |      |
|     |                                 |        |       |            |      |
|     | Jessica                         |        |       |            | ---  |
+  -5 +                                 +        +       +            +(0)   +
-------------------------------------------------------------------------------
|Measr|+suje                                                                   |
```

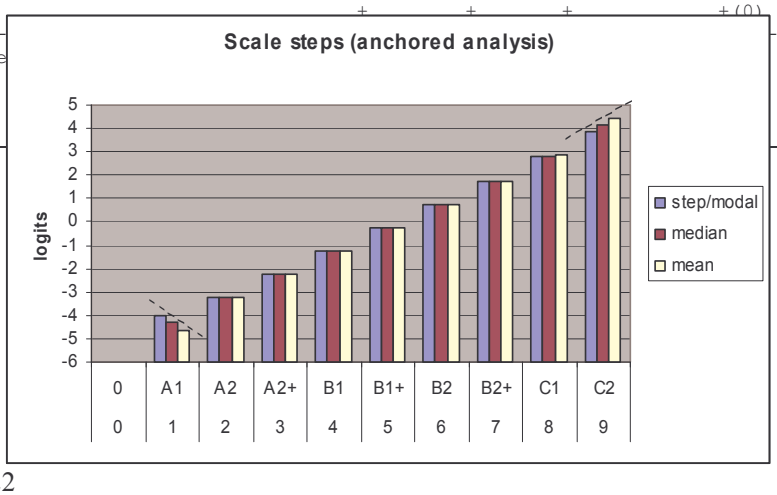


Table 9 Anchored analysis: Levels derived from modal, median and mean cutoffs

		fair average	ability	step/modal	median	mean
33	Josue		9.89	C2	C2	C2
32	Rachel	8.9	6.61	C2	C2	C2
31	Aleksandar DALF	8.1	3.6	C1	C1	C1
15	Ambriogio	7.6	2.97	C1	C1	C1
25	Aleksandar	7.5	2.84	C1	C1	B2+
8	Xi	7	2.24	B2+	B2+	B2+
26	Luis	6.9	2.15	B2+	B2+	B2+
16	Silvia	6.7	1.96	B2+	B2+	B2+
7	Nataliya	6	1.28	B2	B2	B2
37	Gu Jung	5.1	0.34	B1+	B1+	B1+
4	Sophie	5.1	0.3	B1+	B1+	B1+
3	Valérie	4.6	-0.16	B1+	B1+	B1+
13	Evelyne	4.5	-0.28	B1	B1	B1
1	Margarida	4.2	-0.53	B1	B1	B1
5	Debora	3.9	-0.84	B1	B1	B1
2	Mariana	3.8	-0.97	B1	B1	B1
14	Andrea	3.7	-1.02	B1	B1	B1
61	Katell	3.4	-1.37	A2+	A2+	A2+
38	Aamer	3.3	-1.38	A2+	A2+	A2+
62	Sun Ying	3.3	-1.48	A2+	A2+	A2+
6	Iryna	3	-1.73	A2+	A2+	A2+
50	Viggo	1.4	-3.34	A1	A1	A1
42	Suzanne	0.7	-4.18	0	A1	A1
41	Sally	0.7	-4.2	0	A1	A1
49	Jessica	0.5	-4.65	0	0	0

How best to pick among these alternative interpretations? The modal scale is at first sight attractive for two reasons:

1. its meaning – the range where a given rating is the most probable, or frequent – makes intuitive sense;
2. it is what is specified when scale steps are anchored, i.e. it reflects an intention that subjects falling within specified logit ranges should be assigned to certain levels.

However, it is important to understand how anchoring thresholds works. A subject is estimated to be above or below a given threshold irrespective of whether or where that threshold is anchored. Moving the threshold simply causes the subjects move with it.

There appear to be other problems with the modal scale. The nine subjects in the unanchored analysis identified as B1 have fair average scores ranging from 4.6 to 3.2, which is hard to square with the notion of B1 (4) being the most probable or frequent rating over this range. In the anchored analysis the fair average of .7 scored by Suzanne and Sally reflects an actual modal rating of 1 (A1), which seems a more reasonable outcome. In the present analysis it is the median which provides the most credible levels, and it is the scale which is recommended for use in similar situations.

4.2.2 Disattenuated correlation

Correlations between ratings on different performance criteria will be lower if distinct aspects of performance are picked up under each criterion (assuming that the subjects also have distinct profiles of skill). The effect of less-than-perfect reliability also lowers the strength of a correlation. This can be removed by disattenuating the correlations using the formula:

$$r_{disattenuated} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

i.e. the correlation divided by the square root of the reliabilities of both tests multiplied. If a disattenuated correlation is still less than 1 then this can be taken as evidence that the correlated traits are measuring distinct skills.

4.3 Example FACETS data file

```
Title=Sevres Jan 25/05, minus 6 raters, NO anchor to cef levels
output=CEFancNO.out
convergence=200
xtreme=0.5,0.5
facets=6
noncenter=1;
convergence= .1, .001
iterations=250
Vertical = (1*,2*,3N,4N,5*,6N)
Models=
?,18,,?,1,,M,1, ; CEFR treat these raters as missing
?,22,,?,1,,M,1, ; CEFR
?,3,,?,1,,M,1, ; CEFR
?,8,,?,1,,M,1, ; CEFR
?,24,,?,1,,M,1, ; CEFR
?,30,,?,1,,M,1, ; CEFR
?,?,,?,1,,RS1,1, ; CEFR
*
Rating scale=RS1,R10,G,O ; CEFR
 0=0,0,,
 1=A1, -4.01,
 2=A2,-3.23,
 3=A2+,-2.21,
 4=B1,-1.23,
 5=B1+,-.26,
 6=B2,.72,
 7=B2+,1.74,
 8=C1,2.8,
 9=C2,3.9,
*

labels=
1,sujets
1 = Margarida
2 = Mariana
 ... (subjects omitted) ...
61 = Katell
62 = Sun Ying
*

2,raters
1=Christine Tagliante
2=Sylvie Lepage
; ... (raters omitted) ...
37=Sybille Bolton
38=Alba Pardina
39=Neil Jones
*

3,Rater Group
1 = Institutions d'évaluation de langue française
2 = Écoles de langue française (en France)
3 = Projets européens et experts de langue française
4 = Projets européens: autres participants
*

4,Scale
1 = Etendue
2 = Correction
```

```
3 = Aisance
4 = Interaction
5 = Cohérence
6 = N_Global
*

5,Independent
1=oui
2=non
*

6,Jour
1 = Jeudi
2 = Vendredi
3 = Samedi
*

data
e.g. subject 2 rated by rater 37 from rater group 4 on category 1, rating
mode 1, day 1 is rated 4 (B1).
2,37,4,1,1,1,4
2,21,3,1,1,1,3
2,28,3,1,1,1,4
2,23,3,1,1,1,4
2,24,3,1,1,1,2
2,25,3,1,1,1,3
2,26,3,1,1,1,3
 ... (data omitted) ...
```

5 References

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, 37-46,

Linacre J.M. (1993). Rasch-based Generalizability Theory. In *Rasch Measurement Transactions*, 7:1 p.283, retrievable from www.rasch.org

North B, (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York, Peter Laing.

Thompson WD. Walter SD. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*. 41(10):949-58,.

Uebersax, J. (2002a) *Raw agreement indices*. Web page retrieved from http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm

Uebersax, J. (2002b) *Statistical methods for rater agreement*. Web page retrieved from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm