September 2007

Seminar to calibrate examples of spoken performance
Università per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica)
Perugia, 17th – 18th December 2005.

Report on the analysis of the rating data

Michael Corrigan
ALTE Validation Project Co-ordinator

Contents

Introduction

1 aims and considerations on the data

This report presents the analysis of the data from a seminar held in Perugia in December 2005 and hosted by CVCL (Centro per la Valutazione e la Certificazione Linguistica dell'Università per Stranieri di Perugia) in collaboration with the Council of Europe. The principal aim of the seminar was to calibrate samples of spoken Italian to the *Common European Framework of Reference for Languages* (*CEFR*) (Council of Europe:2001) in order for them to be released generally and used for illustrative purposes. This report is principally intended to complement the report on the organisation and proceedings of the seminar itself (Grego Bolli 2006) by illuminating further some of the processes and outcomes of the event. However, the Perugia seminar follows a similar seminar in Sèvres for French (North & Lepage:2005) and one in Munich for German (Bolton:2006) and may precede others. This report will therefore also take account of quantitative studies of the earlier seminars (Jones 2005, 2006), making comparisons with the intention of adding to the pool of knowledge on such events. The selection of contents for this report was based on that in previous reports for comparisons sake and on additional aspects which were salient to the conference organisers.

As described by Grego Bolli (2006), raters were asked to view pre-recorded spoken performances of learners of Italian as a foreign language and rate performances according to the scales of the *CEFR*, using the descriptors provided. All votes were captured electronically in real time. The votes cast can be divided into three separate categories: i) votes on each of five criteria (*range*, *accuracy*, *fluency*, *interaction* and *coherence* – see Council of Europe (2001:ch3)) before discussions of the performance, ii) a *global* vote (representing a holistic consideration of the performance) before discussions and iii) *global* votes after discussions and other related activities (see Grego Bolli (2006) for a description of the activities involved).

Some further explanation about the importance of the differences between the three types of vote may be required. It will be assumed that the difference between votes on rating criteria and *global* votes is clear and does not need further discussion here, except to say that only pre-discussion criteria votes were cast as this was thought to be the most useful methodology (see North & Lepage (2005:16); Jones (2006:14)). The significance of the distinction between votes cast before and votes cast after discussions, however, may require a brief elaboration. The distinction is important because the discussion is expected to have a major effect on the ratings: the views of some participants, along with other activities such as viewing the histogram (see Grego Bolli (2006:5-6)), will have a common influence on opinions and therefore votes and cause there to be less variation. Post discussion votes are therefore more likely to represent greater consensus than pre-discussion votes and this was something intended by the organisers of the first seminar, who aimed to establish 'a consensus in the interpretation of the *CEFR* levels in relation to learner performances in French as a foreign language' (North and Lepage: 2005:3). It should also be added that this desire for a consensus makes sense, given the nature of this event, which aims to produce illustrative samples to be used in many contexts across Europe. However, in terms of the possibilities for analysis, this type of influence can be disadvantageous as is discussed below.

The collection of pre-discussion data is not without its own benefits, however: i) a comparison of two sets of data (pre- and post discussion) can provide information on the effects of the discussion and, ii) the data collected before the discussion allows the application of *Many-Facet Rasch Measurement* (*MFRM*) techniques (see Bachman (2004:146-9) or Bond & Fox (2001:Ch 8) for a more detailed explanation), which require that data are locally independent (i.e. not subject to external influences; see Embretson & Reise (2000:Ch9) for a discussion of this). *MFRM* is especially useful in the rating of performances as, among other advantages, it factors out the leniency/harshness of raters, so the achievement of exact agreement among raters need not be

considered a principal aim in certain rating exercises (Lumley & McNamara: 1995:56; Weigle: 1998:264)[1]. The collection of these two sets of data leaves something of a conundrum, however: post discussion votes are the most meaningful because of greater consensus but the most meaningful analysis (*MFRM*) is only possible on pre-discussion data[2]. For the purposes of the seminar organisers, it was therefore necessary to consider the limitations of each set of data when balancing any conclusions or decisions based on them. For the purposes of this report, the pre-discussion data frequently offered more fruitful opportunities for investigation, as the aim was to provide a quantitative rather than qualitative analysis and, in addition, it was not intended to produce a definitive statement on the nature of the seminar, but rather an exploration into its workings. The results of analysis on pre-discussion data should be seen, therefore, as a snapshot of the event as it evolved towards greater consensus.

2 selection of definitive levels for each performance

The selection of definitive levels for each performance was not a straightforward, mechanical matter. As mentioned above, neither pre- nor post discussion ratings provide the perfect source for a definitive rating. Judgement was required to balance the information available to arrive at the best rating (see Jones (2005:12-14)). Following the precedent of earlier rating events of this kind, it was left to the organising committee to consider the information provided by this analysis alongside other qualitative information gathered during the seminar and allot definitive levels to each performance. This procedure resulted in only one rating which differed from the modal ratings after discussion[3].

Appendix I contains a table which summarises the principal sources of quantitative data used by the organising committee to make their decisions: aggregates of the ratings given to each performance. The first set of columns displays ratings before the discussion and the second set, ratings after the discussion. The final column gives the definitive *CEFR* level given to each performance. Within the columns containing pre-discussion ratings, the first contains the modal (most common) rating and the second contains the result suggested by *MFRM* (i.e. with compensation for lenient or harsh raters). The layout is the same for the post discussion columns but, as mentioned above, the *MFRM* is not appropriate for the analysis of data lacking local independence and, therefore, the second column simply reproduces the modal values, so both post discussion columns are, in fact, the same.

Exploratory analysis

3 levels of agreement

As mentioned above (section 1), a consensus was considered important due to the nature and aims of the event. This section will describe some aspects of this consensus obtained by examining levels of agreement. Facets (Linacre:2005), the software used to conduct the *MFRM* in this analysis, is additionally able to measure the level of agreement between raters. This is defined as the percentage of all ratings made under 'identical conditions [which are] in exact agreement' (Linacre (2006:80-1)). This method is different from that used by Jones (2005, 2006), who adopts raw agreement indices described by Uebersax (2002); both methods, however, are conceptually equivalent.

[1] Consensus is certainly important in the event being discussed here but the aims should not be considered in conflict with contemporary views about performance rating, which suggest that the raters should be encouraged to be self-consistent but not to be consistent with each other (Lumley & McNamara: 1995:56; Weigle: 1998:264), as the context of the benchmarking event is quite different from that of examination grading exercises, where consensus need not be important.

[2] It is possible to run the analysis but at the risk that the data will not fit the model to a degree that it would be difficult to justify any conclusions. That is indeed what happened in this case, where very strong overfit was found and it was therefore decided not to use the *MFRA* of post discussion data.

[3] This performance was that of Desirée, where the pre-discussion vote was adopted. The performance of Marta was rated twice, the second time at the request of the participants. The second rating is recorded as 'Marta Bis'.

3.1 overall levels of agreement

Levels of agreement for this seminar are not dissimilar to those reported by Jones (2005, 2006) and overall agreement is shown in Table 1 ('actual'). Having used Facets (Linacre:2005) to calculate these values, it is also possible to report the levels of agreement expected by the Rasch model ('expected'). The effects of the discussion can be seen when comparing pre- and post discussion levels of agreement. However, a more informative comparison may be found by judging actual levels of agreement in comparison to expected levels of agreement. The effect of the discussion can be seen clearly because the difference between actual and expected agreement is much greater post discussion. As explained in section 1, this implies both greater consensus and less local independence (see section 1).

Table 1 overall levels of agreement (%)

	actual	expected
pre-discussion	40	36.1
post discussion	67.8	56

3.2 levels of agreement by *CEFR* level

Although high levels of agreement were found throughout the data, further investigation can reveal more about the configuration of the agreement in relation to specific parameters. Figure 1 represents the percentage of agreement for performances at each *CEFR* level or each *CEFR* 'plus' level. For the purposes of this part of the analysis, performances were set at a particular level according to the definitive level that they were given as a result of the rating seminar (see Appendix I). Separate bars represent votes cast before or after the discussion. It can be seen that, for each level, post discussion levels of agreement are always higher than pre-discussion levels. It is also clear that agreement is greater at either end of the scale, compared to the middle. This is not dissimilar to the results found in Jones (2005) and Jones (2006) but seems more pronounced here.

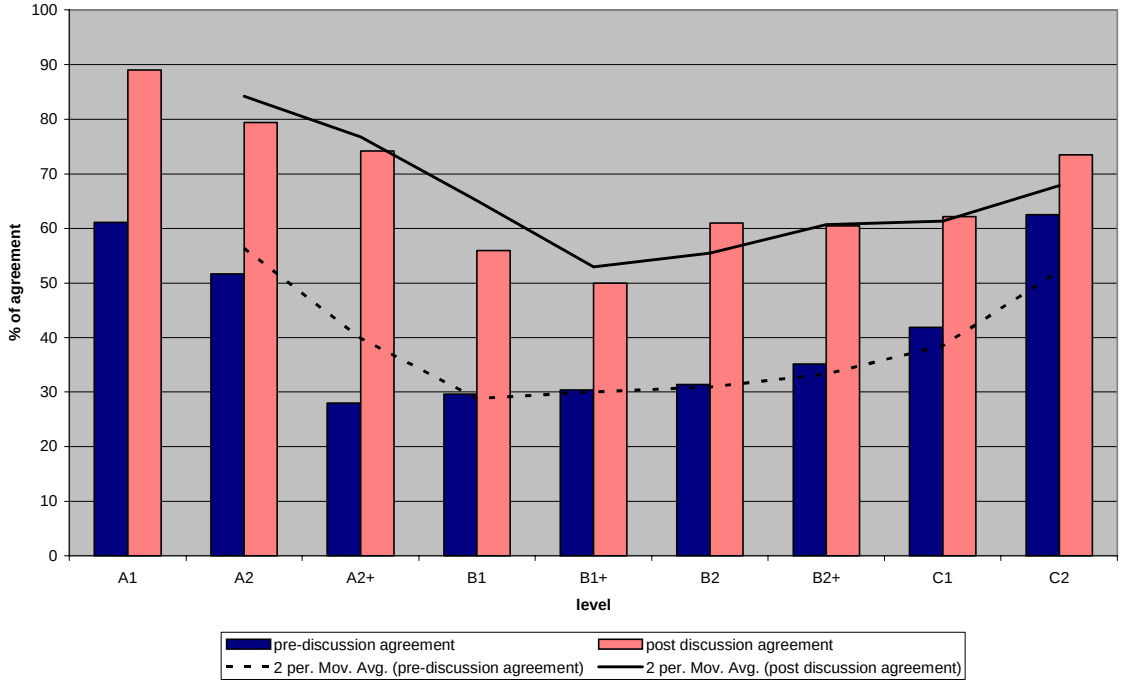


Figure 1 percentage of agreement for performances at rated level

It is possible to investigate the agreement levels displayed in Figure 1 further by considering only votes cast performances which were finally judged to be at the same level. The deviation of votes from the definitive levels are shown graphically in Appendix II. Figure 2 represents a summary of

these charts as it shows all votes as they were cast in relation to the definitive level for each performance. In common with the charts in Appendix II, the frequency of votes is represented by bars placed on the x-axis according to the distance of the votes from the performance's definitive level, set at 0 (e.g. the bar at '-2' represents the all those votes which were cast two levels below the definitive value for each performance). Votes before and after discussion are separated, so that each chart actually contains two distributions, distinguished by colour. It can be seen that both distributions (Figure 2) are very symmetrical and that the distribution of post discussion votes is less spread than that for the pre-discussion votes. This again suggests greater consensus and less local independence.

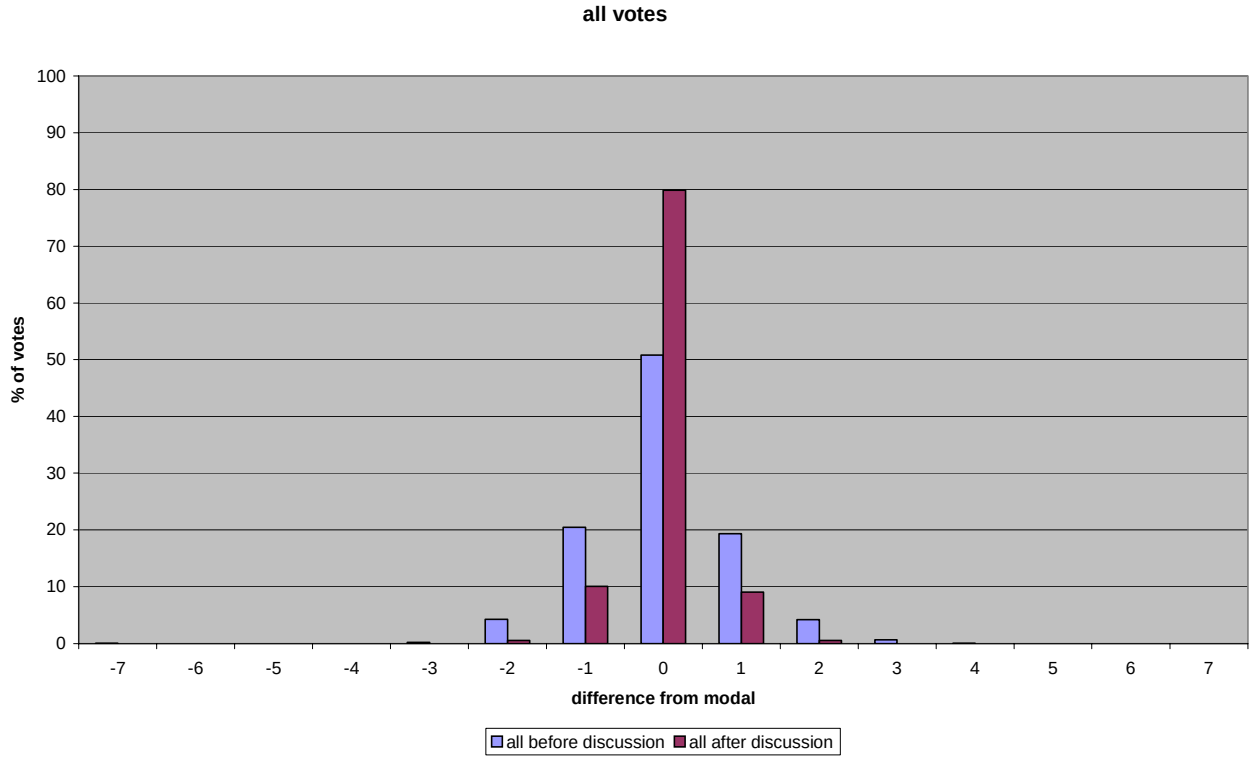


Figure 2 difference of votes from definitive level - all votes, all levels

Figure 3 shows a similar chart to Figure 2 but for a single level: A1. It can be seen that this chart displays less symmetry than Figure 2. The definitive value (at 0) displays no votes to its left-hand side because A1 is at the end of the scale, so there is no choice below this level. Charts for other levels can be seen in Appendix II. The effect of being at or towards the end of the scale can be seen in the first and last two charts. In the case of the last two charts (C1 and C2), the situation is similar to those of the first two but reversed: there is little or no choice above these levels. At all the other levels, the distributions are more evenly spread. Considered together, this series of charts suggests that agreement levels at the extremes of the scale are boosted by there being fewer alternatives within a reasonable range, which concurs with the impression formed by the organisers that obtaining agreement was relatively more difficult in the middle range of the scale. Corresponding phenomena were found by Jones (2005, 2006). The organisers also commented on the difficulty of dealing with certain criteria during voting as sometimes descriptors were not thought to be comprehensive enough. For example, in the absence of guidance from *CEFR* descriptors, participants were unsure at which level the *accuracy* criterion should become a 'truly discriminating criterion'. Whether or not it was intended that descriptors be applied in this way, a reduced amount of agreement over those levels where this was found to be an issue (B1+ to B2+ in particular) may have resulted (also see 4.2).

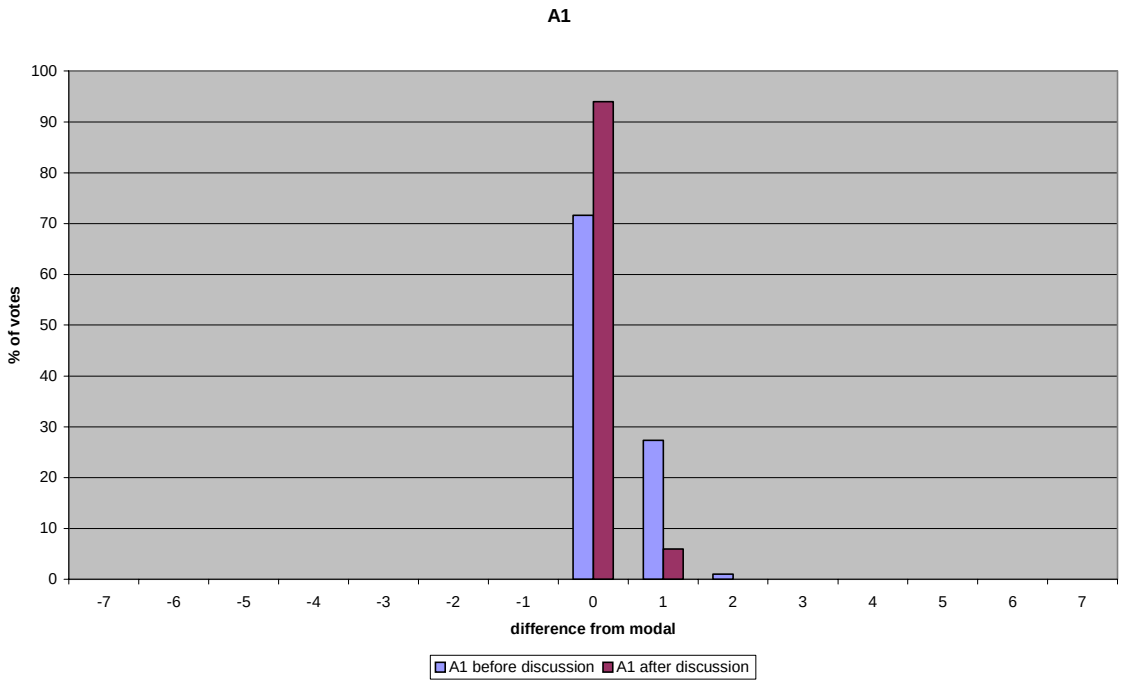


Figure 3 difference of votes from final level - A1

3.3 levels of agreement by rating criteria

Levels of agreement by rating criteria were calculated along with expected levels and are shown in Figure 4. Actual agreement ranges between 36.6% and 41.4%, which is not unlike the 35% to 41% reported by Jones (2005). In the present analysis, the criteria do come in a different rank order, however, but when dealing with such small differences, this is likely to be of little consequence: Jones (2006) reports all figures close to 36%.

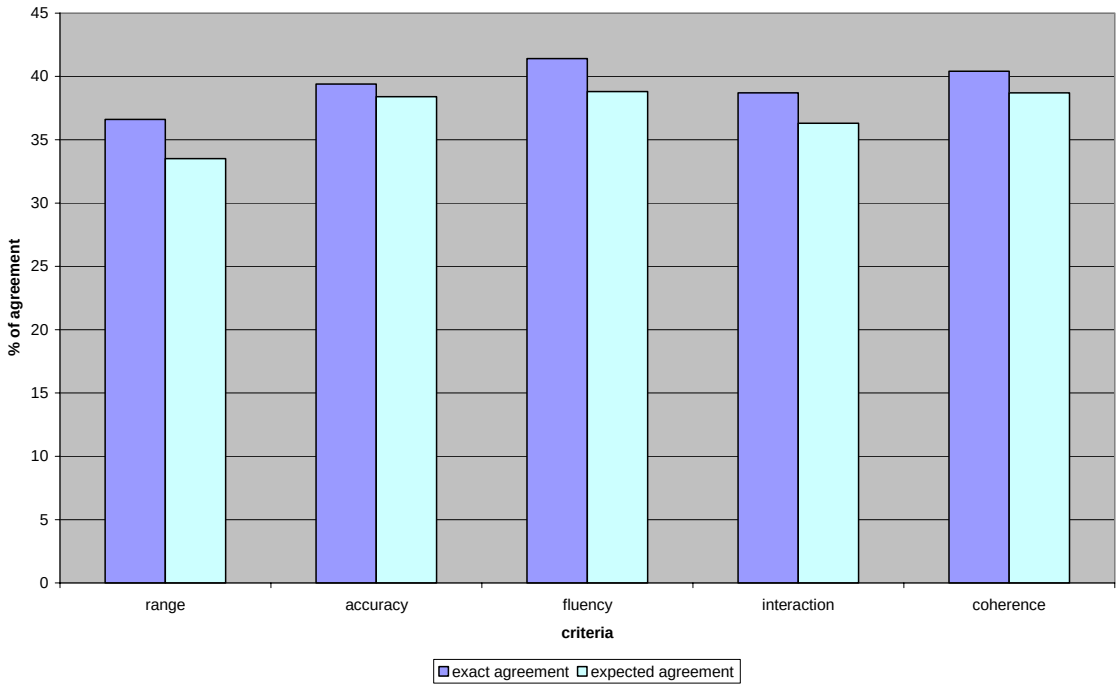


Figure 4 percentage of agreement by criteria

3.4 levels of agreement by day of rating

According to North & Lepage (2005:10), where rating consistency can be measured over time, an improvement should be evident because the activities participants engage in at the event also serve

as training. North & Lepage (2005:10) term this *training effect*. Partly as a consequence, levels of agreement are expected to improve over the duration of such an event (Jones:2005:10, 2006:12). However, unlike previous rating seminars of the same type (Bolton (2006); North & Lepage (2005)), the Perugia seminar was held over two, rather than three days, so a less marked improvement may be expected. In common with previous events (Jones:2005; Jones:2006), it was thought that the best way to examine the *training effect* was to compare votes by the day on which they were cast. Although not investigated directly here, *training effect* may lead to what Myford & Wolfe (2004a:484-5) call *order effects*, where the sequence in which performances are rated is shown to have some kind of influence on the actual rating.

To investigate the *training effect* and *order effects* further, it is necessary to compare the extent to which agreement improved over the duration of the event. However, this is complicated by the fact that the event was organised so as to include a series of plenary discussions throughout, and there is a danger of confounding the effect of the discussions on agreement relating only to a particular performance with the *training effect*, which relates more to underlying rating competence. For this reason, pre- and post discussion data are treated separately. It was again thought important to relate the actual levels of agreement with expected levels of agreement, rather than to compare actual levels of agreement directly, since expected levels of agreement provide a reference with which to judge actual levels. To enable both comparison by day and by vote type, the data was divided into four groups: day one pre-discussion, day one post discussion, day two pre-discussion, day two post discussion; expected levels of agreement were calculated in addition to actual levels of agreement. For each of the four data groups, the difference between actual and expected agreement was then expressed as a percentage of expected agreement. The results are expressed graphically in Figure 5.

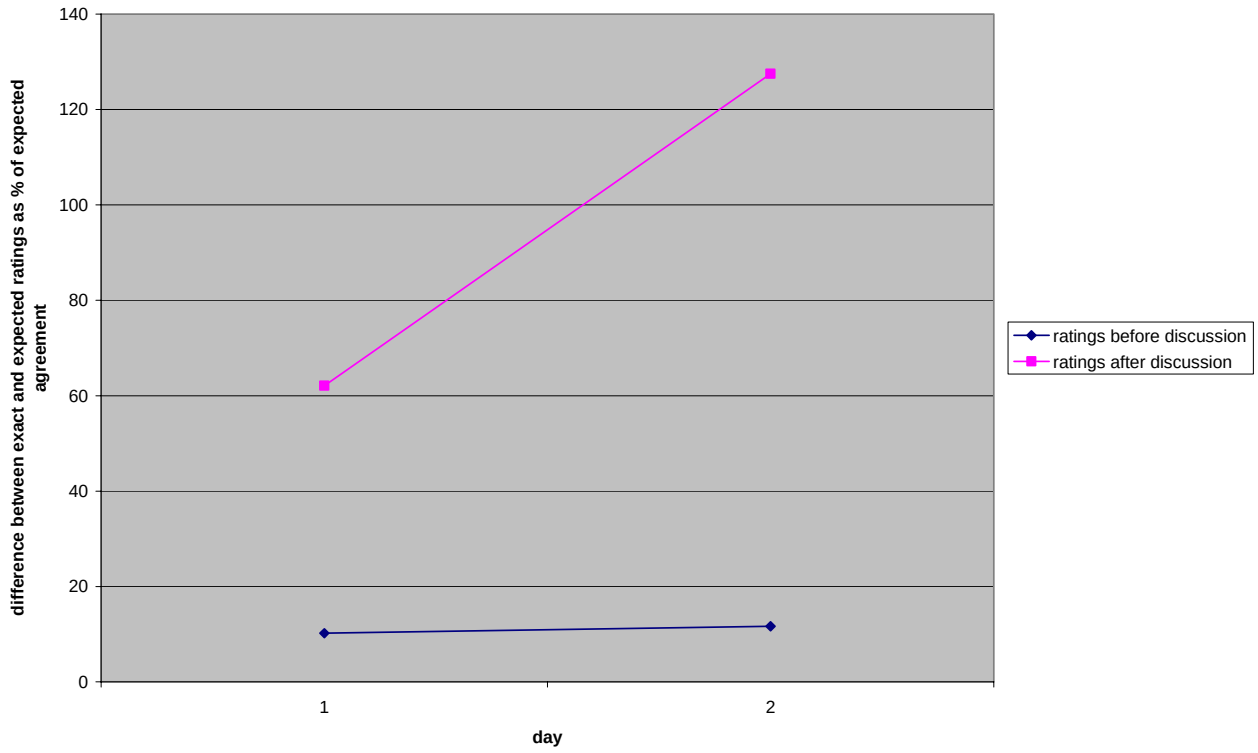


Figure 5 agreement levels by day

It can be seen that i) for each day and each voting type actual agreement is higher than expected agreement (there are no negative points on the chart), ii) there are higher levels of agreement on day two than on day one for both pre- and post discussion votes[4] and iii) agreement for ratings after

[4] The change in pre-discussion agreement over the two days is rather difficult to see. The figures for the chart are: 10.197 on day one and 11.672 on day two.

discussion is much higher than pre-discussion agreement. This final point is the most marked and suggests that whatever *training effect* the activities of the event have on underlying rating competence, agreement still needs to be worked towards for each new sample. The increase in post discussion agreement over two days is also rather substantial (compared to the pre-discussion change), although less so than the difference between pre- and post discussion agreement on the same day. It appears that the increased agreement for day two post discussion votes resulted from the *training effects* having greater impact on the activities related to forming a consensus (such as the discussion) than on activities related to interpreting descriptors individually (such as referring to descriptors). It can be further noted that significant *order effects* are unlikely to have resulted because post-discussion levels of agreement are very high throughout (higher than expected).

4 Use of the rating criteria

4.1 overall severity/leniency in use of rating criteria

One of the most interesting points to investigate in such a seminar is the way in which raters engage with the criteria used for rating. *MFRM* provides an estimate of difficulty for each criterion. This serves as an indication of the relative severity with which raters applied each. Differences between the criteria in this respect may reveal more about the way that descriptors of the *CEFR* function when used in rating. The results can be seen graphically (with error bars) in Figure 6 and numerically in Table 2 and are similar to those found by Jones (2005, 2006), with *accuracy* the most severely rated by far and *interaction* and *fluency* the least. There is some overlap of the error bars for *range*, *fluency* and *global*, which shows that it is not possible to reliably distinguish between the values for these three criteria. The repetition of the pattern of a distinction between *accuracy*, *interaction* and the other criteria suggests that the criteria may have a corresponding structure of salience for raters. This in turn may indicate that the aspects of each performance which correspond to each criterion are more or less important when a global rating is being considered. For example, *accuracy* is applied most severely and the implication of guidance being sought for the use of *accuracy* and not the other criteria (see section 3.2) may be that these raters considered it more relevant than other criteria (at least at some levels). Indeed the salient nature of grammar-related criteria in the rating of speaking is also noted by McNamara (1996:220-2) among others, who also offers some tentative explanations of this phenomenon. On the other hand, *range* and *fluency* do not seem to be used in a way which greatly distinguishes them from the holistic view of the performance, which is recorded in *global*. Where a number of criteria are used in much the same way, what is referred to by North & Lepage (2005:10) and by Myford & Wolfe (2004a:474) as *halo effect* is in evidence.

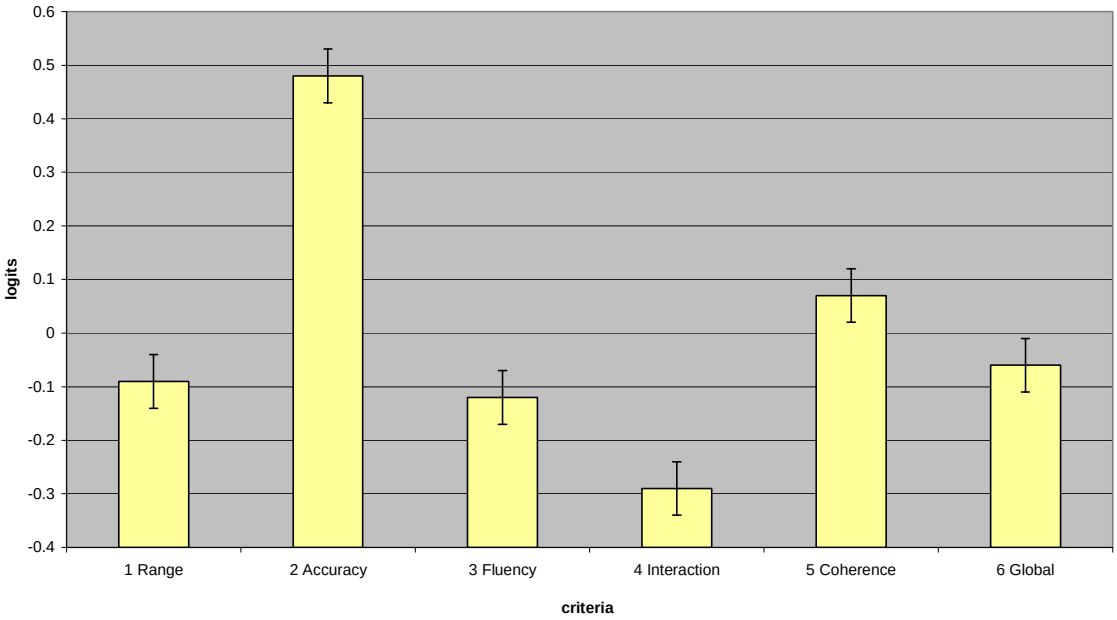


Figure 6 relative difficulty of rating criteria (with error bars)

Table 2 relative difficulty of rating criteria and infit statistics

Criteria	Difficulty	Infit Mn Sq	Infit ZStd
range	-0.09	0.95	-0.70
accuracy	0.48	0.95	-0.70
fluency	-0.12	0.85	-2.30
interaction	-0.29	0.87	-2.00
coherence	0.07	0.81	-3.10
global	-0.06	0.69	-5.40
mean (ex global)	0.01	0.89	-1.76
sd (ex global)	0.29	0.06	1.05

Criteria separation index: 6.72; separation reliability: 0.96; fixed chi-square p = 0.00

4.2 severity/leniency in use of rating criteria by level

Use of the rating criteria can be investigated further by considering whether the severity of criteria ratings differs by the proficiency level of the performance. One of the difficulties reported by participants was to balance the criteria to come to a global rating for each sample (see 4.1). This is particularly challenging when dealing with samples displaying differing ability levels across the criteria and is further complicated because the criteria may vary in salience to raters at different levels (see 3.2 and 4.1).

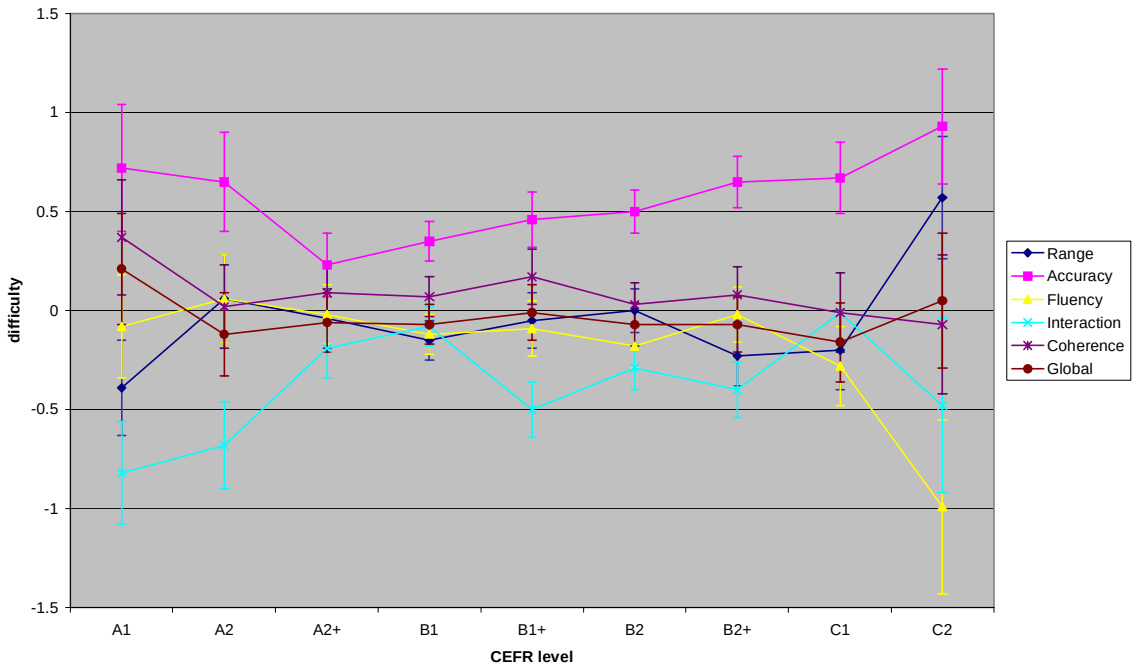


Figure 7 severity of criteria judgements at each *CEFR* level

Performances were grouped according to the final level allotted them (see Appendix I) and, for each group, difficulty estimates for each criteria were calculated from the pre-discussion votes. The results of this analysis are shown in Figure 7, with the *global* rating before discussion also shown as a reference. Error bars have been added, as these show whether the difference in severity of the criteria at each level can be considered separable from each other for the purposes of this analysis. It can be seen that, as with overall criteria severity, where the criteria are sufficiently separable, *accuracy* is always the most severely rated criterion and *interaction* the least. *Range*, *fluency* and *coherence* match the *global* rating on most of the levels, with overlapping error bars.

Table 3 quality statistics for criteria difficulty estimations by level

	entire population			
	seperation	trait seperation index	reliability	fixed Chi-sq p
A1	1.57	3.43	0.71	0.00
A2	1.50	3.33	0.69	0.00
A2+	0.00	1.33	0.00	0.47
B1	1.40	3.20	0.66	0.00
B1+	1.76	3.68	0.76	0.00
B2	2.05	4.07	0.81	0.00
B2+	2.11	4.15	0.82	0.00
C1	1.25	3.00	0.61	0.00
C2	1.47	3.29	0.68	0.00

The extent of what appears to be *halo effect* at each level can be seen easily in numerical terms by viewing the quality statistics in Table 3 which are arranged by level. It can be seen that the trait separations index, which 'connotes the number of statistically distinct levels of trait difficulty among the traits included in the analysis' (Myford & Woolfe:2004b:549), is always less than the number of traits (criteria) in the analysis. This can be contrasted to that given with Table 2, which is 6.72. The A2+ level is the most problematic, showing no clear statistical separation of difficulty levels.

The *halo effect* examined in this report, is, of course, based solely on pre-discussion data. It remains unclear what differences there may have been had criteria votes been cast post discussion and

fruitful way of analysing them found. However, the difference between pre- and post discussion agreement levels described in section 3.4 should be remembered, as it seems to suggest that raters are far more decisive after the discussion. In addition, it should be noted that the use of criteria descriptors may have important training effects.

4.3 consistency in use of the rating criteria

Another aspect on which to compare the use of rating criteria is the consistency with which each is applied. This can be investigated by reviewing the criteria *Infit ZStd* statistics (Table 2) produced through *MFRA*, as what they summarise can be thought of as a measure of the predictability of each vote (Linacre:2002). The statistic is based on the difference between an individual vote and what is expected by the Rasch model for that vote, given the severity of the rater, the difficulty of the criteria, the ability displayed in the performance and so on for other parameters. Greater variation from what is expected indicates less consistency on the part of the rater. The *Infit ZStd* statistic has an expected value of 0, with positive and negative values indicating more of less consistency. A value of less than -2 is usually considered overly consistent; a value of more than +2 is usually considered too inconsistent.

Of the values reported by Jones (2005:9, 2006:11) for the two previous seminars, the present one seems closer to that in Sèvres. The range between the highest and lowest value is around the same, the most overfitting (most consistently applied) criterion is *coherence* and the least is *accuracy* (jointly with *range* in the case of the Perugia). All Perugia values, however, display overfit (more consistency than expected), whereas not all the Sèvres values do. This means that, despite reports of difficulty in using the *accuracy* criterion, it was applied in a more consistent fashion than expected. *Range* does not strongly overfit here as it does in the Munich data. Table 2 also contains figures for the pre-discussion *global* votes, which overfit more strongly than any of the criteria. This may be partly because voting on a single, holistic criteria is somehow inherently easier and/or because the *global* vote always followed consideration of the individual criteria, which influenced raters' ideas about performances (see 3.4 and 4.2).

4.4 generalisability of the rating criteria

Although performances were rated on five criteria, it could be possible to ask how far the ratings could be generalised to situations where fewer criteria were used (i.e. would the results be the same were fewer criteria used to rate performances ?). Generalizability theory (see Cronbach (1990:195-7) for a fuller explanation) provides a way in which this can be estimated. For this analysis, the pre-discussion ratings for 24 performances by 21 raters were included to ensure a balanced design, with no gaps, and the data was analysed using GENOVA (Crick & Brennan:1984). In the analysis, 88.239% of variance in the data was found to be due to performances. Table 4 displays the G-coefficients and phi coefficients for situations where five criteria are used and where just one criterion is used. The values are very high in both cases. This is not surprising as *global* votes were cast after the criteria votes and did not show any unusual variation from the criteria votes. The generalisability of the number of raters will be discussed below (5.3).

Table 4 generalizability and phi coefficients for different numbers of criteria and raters

	number of rating criteria			
	5		1	
number of raters	G-coefficient	phi	G-coefficient	phi
35	0.99795	0.99779	0.99661	0.99617
2	0.96527	0.96267	0.9438	0.93703
1	0.93286	0.92803	0.89358	0.88153

In concord with much of the rest of section 4, the high coefficients yielded by the generalizability study also show that, despite some differences, overall, there is a great amount of similarity in the way that criteria are employed, which suggests a *halo effect*. For the sake of comparison with earlier

studies (Jones:2005; Jones:2006), further evidence of the *halo effect* can be seen when examining the correlation between the rating criteria (Myford & Wolfe:2004a:474). If the criteria were used independently, they would be unlikely to correlate highly with one another. High correlation coefficients can be seen in Table 5, with the results here very similar to those in both other studies (Jones:2005:8, 2006:9).

Table 5 correlation of independently estimated abilities for each rating criteria

1 range	1.000	0.988	0.992	0.987	0.993
2 accuracy		1.000	0.996	0.981	0.997
3 fluency			1.000	0.989	0.997
4 interaction				1.000	0.986
5 coherence					1.000
	1 range	2 accuracy	3 fluency	4 interaction	5 coherence

correlation significant at the 0.02 level (two-tailed)

4.5 rating criteria profiling

Jones (2006:9-10) displays charts which show an ability profile for individual performances across the five criteria. This was done with profiles where at least one significant bias effect was evident by subtracting the bias estimate for each performance from the overall ability rating for that performance. The results seemed to suggest that performances were rated more or less severely than expected on particular criteria in order to conform to the overall assessment of the performance. The pattern that suggested itself was such that the *accuracy* and *interaction* criteria were rated at opposite extremes, so that a performance rated high on *accuracy* was rated low on *interaction* to compensate, and vice versa. As only one significant bias effect was in evidence in the current data, it was not possible to investigate this phenomenon further.

4.6 use of descriptors

Previously in this report, it has been noted that raters found some difficulties in using the descriptors provided (sections 3.2 and 4.2). Difficulties in applying descriptors are likely to lead to problems such as inconsistent ratings, which in turn produce unclear distinctions when ratings are aggregated for the purposes of establishing definitive ratings. As *CEFR* levels form a sequential system, with each successive level representing a higher level of ability than the last, lack of clarity in one part of the scale affects the integrity of the whole scale. With *MRFA*, a rating scale can be represented as the likelihood of any rating being chosen by raters for a performance of a particular ability level. In a good rating scale, each level is the most likely for a range of abilities, the range for each level is approximately the same length as that for other levels, the likelihood in each case is approximately of the same magnitude and these levels are ordered in the sequence intended.

The graphical representation of the rating scale constructed from the pre-discussion votes can be seen in Figure 8; the numerical representation is in Table 6. It can be seen in the graphical representation, for example, that a performance displaying an ability of -7.6, is most likely to attract the rating of A1 because the blue line representing A1 is higher than any other curves at this point. This also true for any point within the whole range where the blue A1 curve is higher than the curves for other levels (-9.92 to -3.13). The *step calibrations* that mark either end of the range where each level is most likely are displayed in Table 6 and are, in addition, the points at which two levels are equally likely to be given to a performance; the size of this range is also given.

Table 6 Counts and step calibrations for rating scales

Original data								
Response Category Name	count	%	STEP CALIBRATIONS Measure	size	(logits) S.E.	QUALITY Average Measure	Exp. Measure	OUTFIT MnSq
A0	1	***0%***				2.55	-4.44	5.5
A1	363	11%	-9.92	6.79	0.99	***-3.74***	-3.73	0.9
A2	457	14%	-3.13	2.53	0.08	-2.01	-1.98	0.9
A2+	261	8%	-0.6	0.07	0.08	-0.62	-0.55	0.8
B1	371	11%	-0.53	1.21	0.07	0.19	0.17	0.7
B1+	337	10%	0.68	0.69	0.07	0.94	1.04	0.9
B2	422	13%	1.37	1.22	0.07	2.16	2.13	0.8
B2+	429	13%	2.59	0.86	0.07	3.08	3.02	0.7
C1	445	13%	3.45	2.64	0.07	4.22	4.17	0.9
C2	226	**7%**	6.09		0.11	6.7	6.76	1

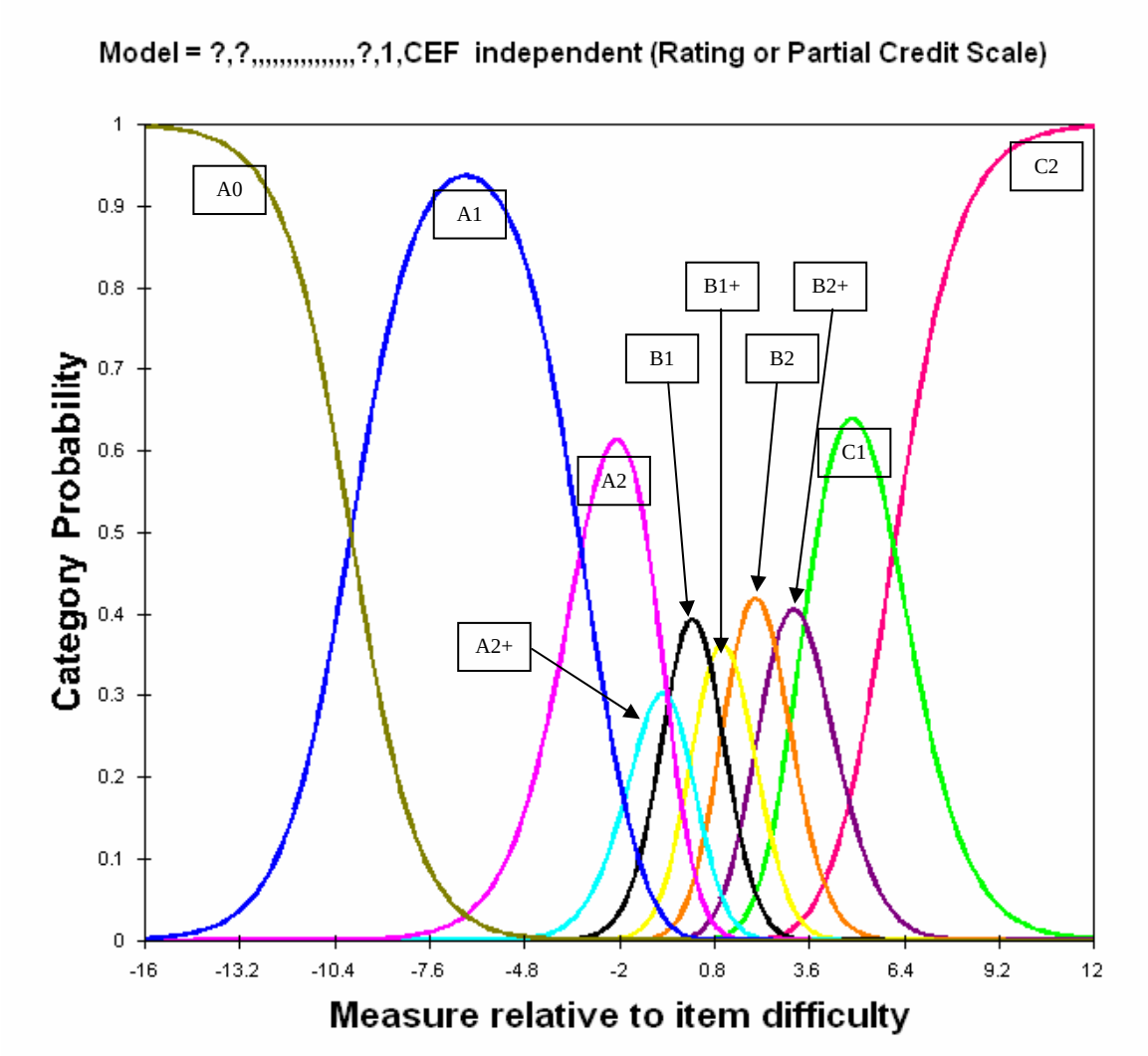


Figure 8 Category curves for 'plus' level rating scale

Four difficulties are immediately obvious from Figure 8 and Table 6: i) one of the categories (A2+) is at no point the most likely rating, ii) the categories are not evenly sized or spread, iii) the measures within categories are not monotonic (sequential – see 'Average Measure' in Table 6), and iv) the numbers of votes in some categories (A0) is unbalanced (see 'count' in Table 6). It was decided that a rationalisation of the scale would be useful in order to investigate the issues related to these outcomes further. In the guidelines presented by Bond and Fox (2001:167) for collapsing rating scale categories, 'the first and foremost guideline…is that what we collapse must make sense'. For this reason, the 'plus' level data were recoded so that they were subsumed by the

standard *CEFR* categories: A0 was collapsed into A1, A2+ into A2, B1+ into B1 and B2+ into B2. The likelihood curves from the analysis of this recoded data can be seen in Figure 9.

In contrast to the original data, scrutiny of Figure 9 and the 'recoded data' part of Table 6 reveals i) each level having a difficulty range where it is the most likely, ii) more evenly sized and evenly placed categories, and iii) monotonic measures within categories. Linacre's (1995) recommendation that frequency counts within categories be as balanced as possible was evidently not possible and the size of the counts in most conflated categories has approximately doubled. However, no specific problems seem to result from this, whereas the problems resulting from under representation in A0 have been dealt with.

Table 7 Counts and step calibrations for recoded rating scales

Recoded data								
Response Category Name	count	%	STEP CALIBRATIONS Measure	size	(logits) S.E.	QUALITY CONTROL Average Measure	Exp. Measure	OUTFIT MnSq
A1 (A0, A1)	364	11%				-7.62	-7.75	1.1
A2 (A2, A2+)	718	22%	-7.01	4.21	0.09	-4.63	-4.51	0.9
B1 (B1, B1+)	708	21%	-2.8	3	0.07	-1.35	-1.28	0.8
B2 (B2, B2+)	851	26%	0.2	3.32	0.07	1.95	1.87	0.8
C1	445	13%	3.52	2.57	0.07	4.03	3.94	0.9
C2	226	7%	6.09		0.11	6.79	6.86	1

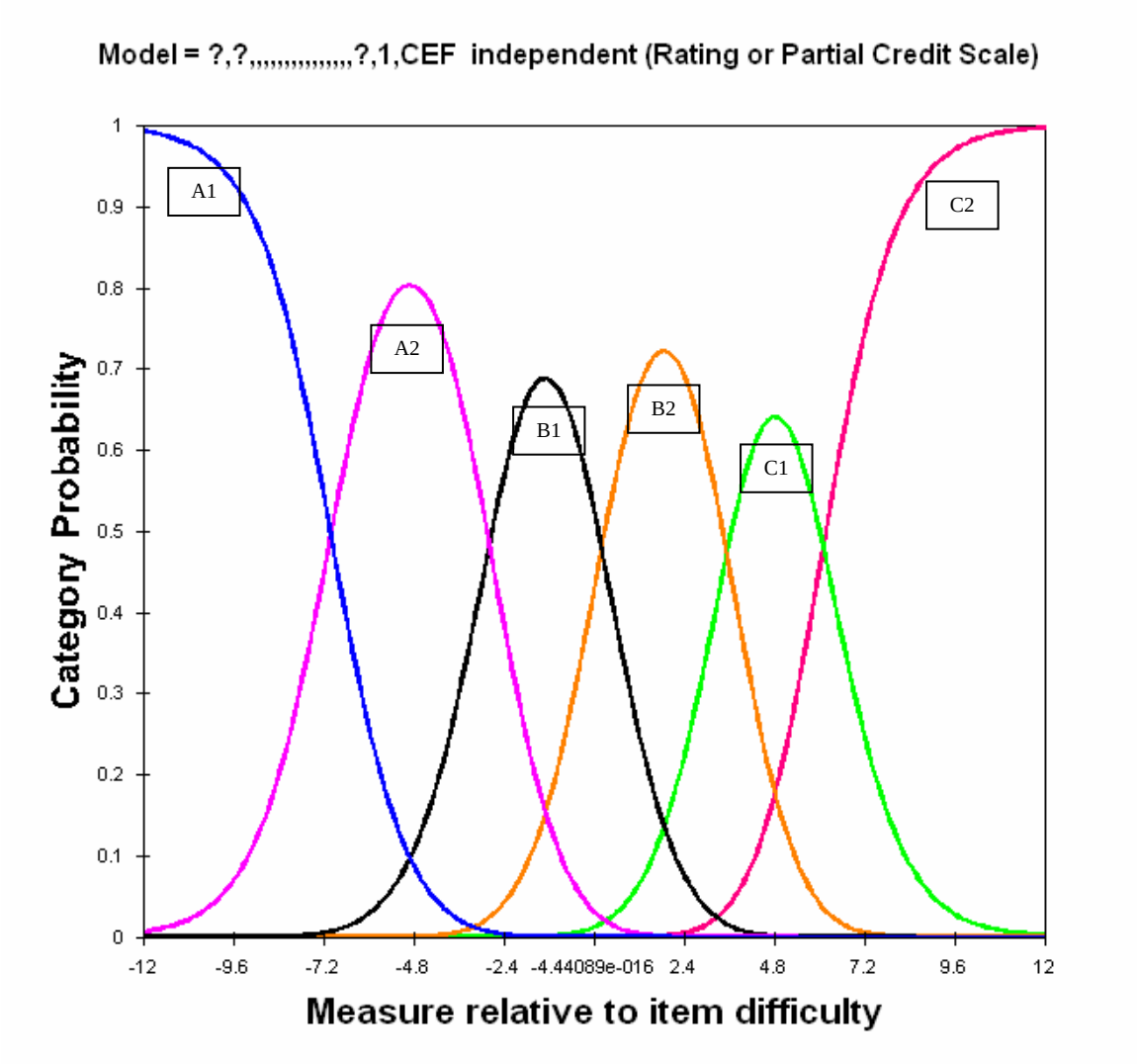


Figure 9 Category curves for *CEFR* rating scale

A number of factors are likely to contribute to the pre-discussion rating scale illustrated in Figure 9. One cause may be that familiarisation with the scales and the process is not just a discrete step which precedes rating but something that continues during the rating itself (as discussed in 3.4). Another cause may be that the scales are not entirely adapted for this specific use. Table 3 of the *CEFR* is described as 'assessor-orientated' (Council of Europe:2001:38) according to Alderson's (1991) distinction. However, North's (2002) account of the inception of the 'plus' level descriptors may indicate that some of the scales are to some extent 'user-orientated' and therefore not fully adapted to use for rating. North (2002:95-100) reports that after the initial analysis of data from the use of the scales, the bands representing the *CEFR* levels were not of an even length on the scale, so some discussion followed over whether to introduce 'plus' levels to reduce the length of the larger bands. On one side of the discussion, the idea was put forward that wider ranges for the bands would allow clearer distinctions when using descriptors; on the other, that narrower bands would allow learners to have a greater sense of progress and therefore increased motivation. It was finally decided to provide 'plus' level descriptors separately for use by those who chose to use them. On another tack, Pollitt (2004:5) puts forward the argument that a maximum of five categories in a scale can be used before the number of categories has a negative impact on raters' ability to use them properly.

5 Rater Behaviour

5.1 rater severity/leniency

Leniency/severity effect, which Myford & Wolfe (2004a:471-3) suggest may be seen as a consistent tendency of the rater to rate more leniently or severely than other raters, is something that *MFRM* is well equipped to detect and measure. Rater difficulty estimates are placed on a scale from negative to positive infinity, with 0 representing raters neither lenient nor severe. The histogram of the distribution is shown in Figure 10 and, by gauging the position of the distribution relative to zero, it can be seen that this group of raters tended slightly towards the severe. One rater[5] (rater 26, at -1 in Figure 10) is clearly far more lenient than the others but with and without this outlier, the spread of this distribution is very similar to that reported by Jones (2005:7): the ratio of the standard deviation of rater severity to the standard deviation of the abilities estimated for performances is 8.5:1 including the outlier and 11:1 excluding the outlier. The danger of *central tendency effect* (overuse of middle rating categories – see Myford & Wolfe:2004b:531) seems small as the performances separation index in the performances table (Appendix III) shows a large degree of separation (34.68). Indeed, there is little reason to think that raters might be inclined towards *central tendency effect* at an event such as this, as they are unlikely to feel pressure to 'play it safe' by using only a small part of the scale (Myford & Woolfe:2004:532), in contrast to what might happen in a monitored grading exercise.

[5] This rater was a non-native speaker of Italian who had no previous experience rating performances in Italian as a foreign language. This rater participated in fewer votes than most other raters.

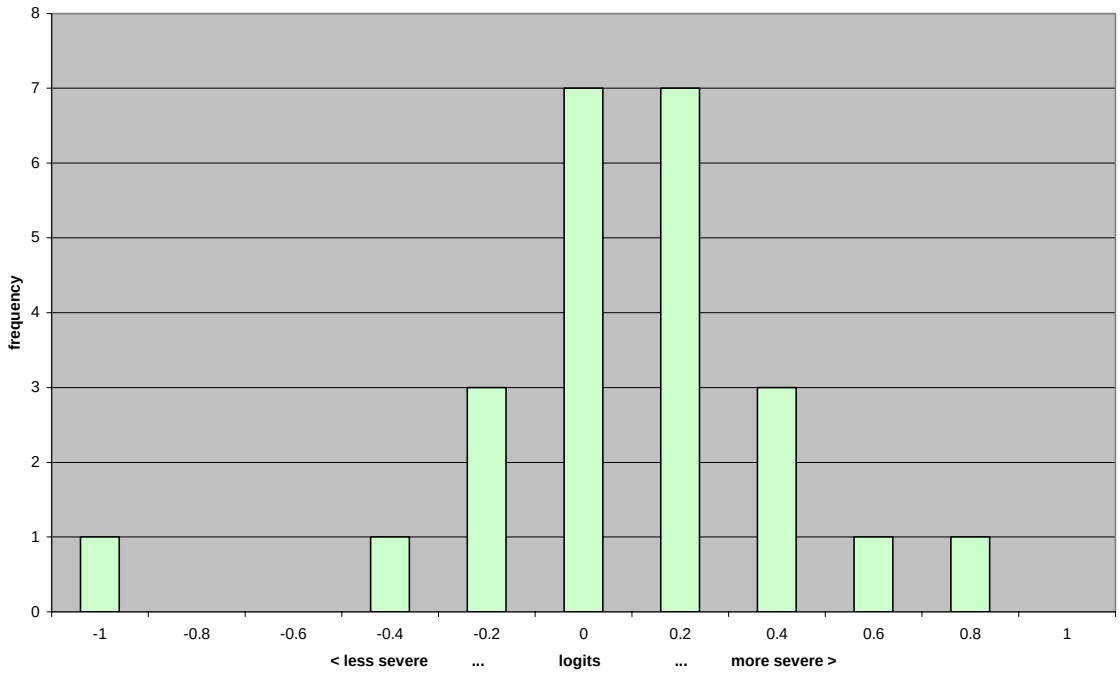


Figure 10 distribution of raters by severity

As in section 4.3, an analysis of fit statistics can shed light on the consistency. The mean Infit ZStd in the 'raters' table (Appendix III) shows that there is some degree of overfit, or greater than expected consistency (the figure is negative). Jones (2006:7) reports something similar and suggests that it may be that the familiarisation phase of the event ensured that none of the votes which followed were truly independent. In any event, the Infit ZStd statistic is not extreme and falls within the acceptable range of -2 to +2.

5.2 the influence of other characteristics on rater behaviour

Myford & Wolfe (2004a:481-2) describe two families of effect that may be expected to apply to groups of raters: i) *influences of rater/ratee background characteristics* and ii) *influences of rater bias, beliefs, attitudes and personality characteristics*. The former relate to demographic characteristics such as native/non-native 'speakerhood', where, for example, non-native speaking raters might be systematically more severe. The latter type of influence appears to more helpfully related to individuals than to groups. However, in the context of this seminar, raters' working background may be said to influence beliefs and attitudes, for example, raters may by their use of more familiar frameworks used on a regular basis, and find it difficult to use the *CEFR* descriptors provided.

An investigation of the possible effects of belonging to a distinct group was made by Jones (2005:11, 2005:13). For the present analysis, groups were identified according to characteristics salient to the seminar organisers Grego Bolli (2006:6-7) and those obvious for other reasons (e.g. gender) (see Appendix IV for a full list of rater groups used). Groups were analysed as facets of difficulty using *MFRM* in an attempt to detect differences between the groups, or bias. In both instances, as with Jones (2005, 2006), nothing significant was found.

5.3 generalisability of ratings

An important element of the performance of raters is that of reliability. This is interpreted by generalizability theory as the extent to which outcomes would be similar were certain elements of the situation changed. In addition to showing figures for the generalizability of the rating criteria, Table 4 also gives coefficients for raters. The coefficients are high and suggest that the results of

this analysis would not be different with other raters, with thirty-five raters, or with a smaller number of raters.

Conclusion

At each point where comparisons of the results of the present analysis and those of the previous analyses (Jones:2005; Jones:2006) were possible, they have been made in the text of this report. Throughout, these comparisons gave little reason to suggest any fundamental differences between the three events in terms of the nature of the processes through which samples were allotted a level:

agreement
overall levels of agreement were high, the pattern of agreement across levels was similar as was agreement at particular levels and for particular criteria. Agreement was higher for post discussion votes and for votes on the second day.

use of criteria
the pattern of severity among the rating criteria and consistency was similar, as was the generalisability of the ratings.

behaviour of raters
the spread of the severity and consistency of raters was similar, there was little influence of other rater characteristics was in evidence and ratings were highly generalisable.

In addition to replication of the work done in previous studies, some new analysis has been undertaken. This analysis included an investigation into the use of the rating criteria at different *CEFR* levels and one into the use of the rating scale. Through these additional investigations and the analysis in the rest of the report, it can be concluded that, at the pre-discussion stage of the event, opinions on the performances were, in general, still embryonic. However, as section 3.4 shows, there is a large difference in the level of agreement between pre- and post discussion ratings, so it may be expected that other aspects of post discussion voting, which it was not possible to analyse in a comparable way, were also quite different. Of particular interest to organisers of similar events in the future may be the considerations in this report of the difference between pre- and post discussion levels of agreement (3.4), use of the rating criteria (4.1 to 4.3) and the use of rating scale (4.6).

References

Alderson, J. Charles (1991) Bands and Scores in J. Charles Alderson & Brian North (eds.) *Language Testing in the 1990s*. London: Macmillan.

Bachman, Lyle F. (2004) *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bolton, Sybille (2006) Seminar to calibrate sample of spoken performances to the Common European Framework of Reference for Languages. Goethe-Institut, Zentrale, Munich, 19th – 22nd October 2005. Report. Retrieved 4th April 2007 from: http://www.coe.int/T/DG4/Portfolio/main_pages/Report%20Seminar%20in%20German.pdf

Bond, Trevor G. & Fox, Christine M. (2001) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah NJ: Lawrence Erlbaum Assoc.

Council of Europe (2001) The Common European Framework of Reference for Languages: learning, teaching, assessment. Retrieved 4th April 2007 from: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Crick, Joe E. & Brennan, Robert L (1984) *GENOVA computer program*. Version 2.2. Iowa City, IA: American College Testing Program.

Cronbach, Lee J. (1990) *Essentials of Psychological Testing (fifth edn.)*. New York: Harper & Row.

Embretson, Susan E. & Reise, Steven P. (2000) *Item Response Theory for Psychologists*. Mahwah NJ: Lawrence Erlbaum Assoc.

Grego Bolli, Giuliana (2006) Seminar to calibrate examples of spoken performances in Italian L2 to the scales of the Common European Framework of Reference for Languages. Università per Stranieri di Perugia – CVCL (Centro per la Valutazione e la Certificazione Linguistica). Report. Retrieved 4th April 2007 from: http://www.coe.int/T/DG4/Portfolio/main_pages/Report%20on%20Italian%20Benchmarking%20seminar.pdf

Jones, N. (2005) Seminar to calibrate examples of spoken performance: CIEP Sèvres, 02-04 December.2004. Report on analysis of rating data. Retrieved 4th April 2007 from: http://www.coe.int/T/DG4/Portfolio/documents/SevresreportNJ.pdf

Jones, N. (2006) Seminar to calibrate examples of spoken performance: Goethe-Institut, Munich, November 2005. Report on analysis of rating data. Unpublished manuscript.

Linacre, J.M. (1995) Categorical Misfit Statistics. Rasch Measurement Transactions, 9/3, 51. Retrieved 16th May 2007 from: http://www.rasch.org/rmt/rmt93j.htm

Linacre, J.M. (2002) What do Infit and Outfit, Mean-square and Standardized mean? Rasch Measurement Transactions, 16/2, 878. Retrieved 16th May 2007 from: http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2006) *A User's Guide to FACETS MINIFAC Rasch-Model Computer Programs*. Retrieved 4th April 2007 from: http://www.winsteps.com/aftp/facets.pdf

Linacre, J. M. (2005) *Facets Rasch measurement computer program*. Version 3.58.0. Chicago: Winsteps.com.

Lumley, Tom & McNamara, T.F. (1995) Rater characteristics and rater bias: implications for training. *Language Testing*, 12, 54 – 71.

McNamara, T.F. (1996) *Measuring Second Language Performance*. Harlow, Essex: Longman.

Myford, Carol M. & Wolfe, Edward W. (2004a) Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I in Everett V. Smith Jr. & Richard M. Smith (eds.) *Introduction to Rasch Measurement: Theory, Models and Applications*. Maple Grove MN: JAM Press.

Myford, Carol M. & Wolfe, Edward W. (2004b) Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part II in Everett V. Smith Jr. & Richard M. Smith (eds.) *Introduction to Rasch Measurement: Theory, Models and Applications*. Maple Grove MN: JAM Press.

North, Brian (2002) Developing descriptor scales of language proficiency for the CEF Common Reference Levels in J. Charles Alderson (ed.) *Common European Framework of Reference for Langauges: Learning, Teaching, Assessment – Case Studies*. Retrieved 4th April 2007 from: http://www.coe.int/T/DG4/Portfolio/documents/case_studies_CEF.doc

North, Brian & Lepage, Sylvie (2005) Seminar to calibrate examples of spoken performances in line with the scales of the Common European Framework of Reference for Languages. CIEP Sèvres, 02-04 December.2004. Report. Retrieved 4th April 2007 from: http://www.coe.int/T/DG4/Portfolio/documents/reportsevres.pdf

Pollitt, Alistair (2004) Let's stop marking exams. Retrieved 4th May 2007 from: http://www.cambridgeassessment.org.uk/research/confproceedingsetc/IAEA2004AP

Uebersax, J. (2002) *Raw agreement indices*. Web page retrieved 4th April 2007 from http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm

Weigle, Sara Cushing (1998) Using FACETS to model rater training effects. *Language Testing*, 15, 263 – 287.

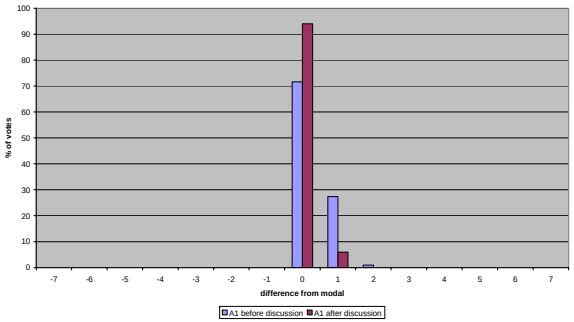
Appendix I: final levels

	judgement before discussion			judgement after discussion			
		raw votes			raw votes		definitive
#	name	modal	Facets	name	modal	Facets	rating
5	RINA	C2	C2	RINA	C2	C2	C2
6	ESTERE	C2	C2	ESTERE	C2	C2	C2
10	ELLI	C1	C1	ELLI	C1	C1	C1
14	KARINE	C1	C1	KARINE	C1	C1	C1
13	KIM	C1	C1	KIM	B2+	B2+	B2+
18	RAQUEL	B2+	B2+	RAQUEL	B2+	B2+	B2+
9	WILMA	B2+	B2+	WILMA	B2+	B2+	B2+
17	AMALIA	B2	B2+	AMALIA	B2	B2	B2
1	MARTA	B2+	B2+	MARTA	B2+	B2+	
				MARTA Bis*	B2	B2	B2
21	DIMITRIO	B2	B2	DIMITRIO	B2	B2	B2
7	AGATA	B2	B2	AGATA	B2	B2	B2
8	SIMON	B1+	B1+	SIMON	B1+	B1+	B1+
22	VERONICA	B1+	B1+	VERONICA	B1+	B1+	B1+
2	MEGUMI	B1	B1	MEGUMI	B1	B1	B1
19	STEFANIE	B1	B1	STEFANIE	B1	B1	B1
20	EWA	B1	B1	EWA	B1	B1	B1
24	DESIREE	B1	B1	DESIREE	A2+	A2+	B1
23	CRAIG	A2	B1	CRAIG	A2+	A2+	A2+
15	SOPHIE	A2+	A2+	SOPHIE	A2+	A2+	A2+
16	MOHAMMAD	A2	A2	MOHAMMAD	A2	A2	A2
12	MALEGIO	A2	A2	MALEGIO	A2	A2	A2
4	ALLISON	A1	A1	ALLISON	A1	A1	A1
3	DIANA	A1	A1	DIANA	A1	A1	A1
11	BRUCE	A1	A1	BRUCE	A1	A1	A1

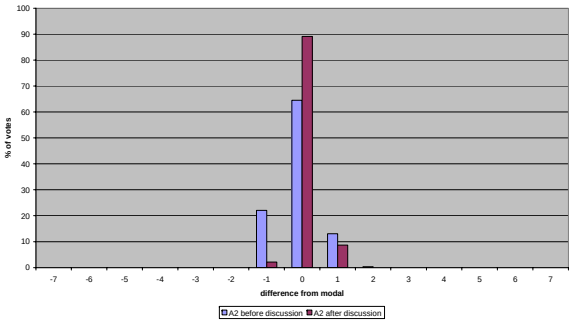
*represents a second rating for Marta

Appendix II: agreement charts

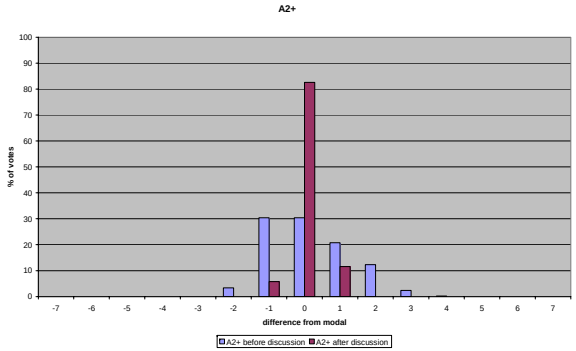
A1



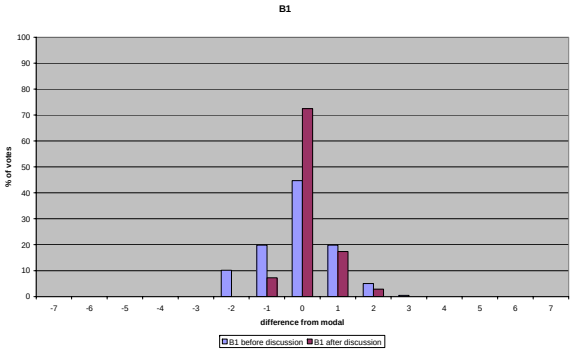
A2



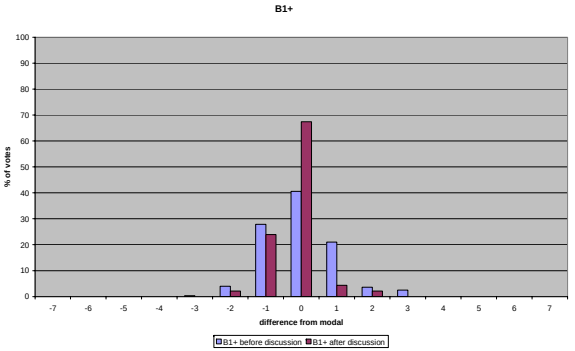
A2+



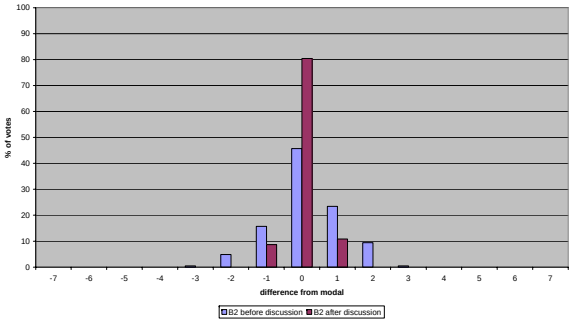
B1



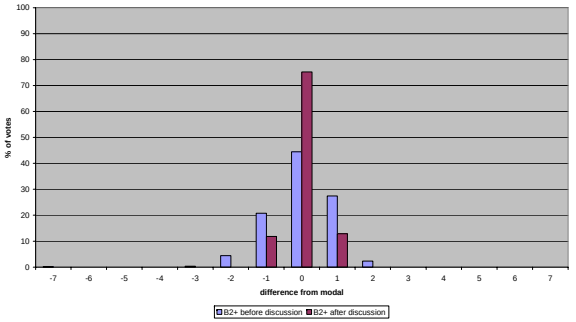
B1+



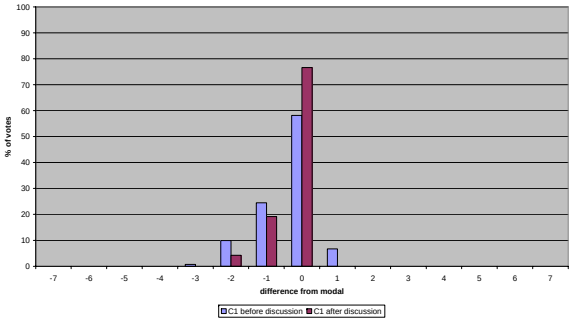
B2



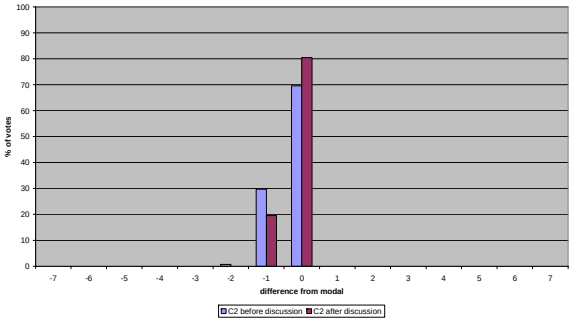
B2+



C1



C2



Appendix III: rater and performance statistics (pre-discussion data)

raters

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Raters
636	144	4.4	4.26	0.68	0.1	0.81	-1.5	0.88	-0.9	1.05	21
646	144	4.5	4.37	0.59	0.1	1.11	0.8	1.02	0.2	0.95	11
668	144	4.6	4.6	0.39	0.1	0.88	-0.9	0.88	-1	1.12	22
680	144	4.7	4.72	0.28	0.1	1.09	0.7	1.08	0.6	0.88	6
687	144	4.8	4.79	0.22	0.1	1	0	1.04	0.3	0.88	9
340	72	4.7	4.81	0.2	0.12	0.59	-2.9	0.56	-3.1	1.53	19
692	144	4.8	4.84	0.17	0.1	0.72	-2.4	0.72	-2.5	1.26	24
692	144	4.8	4.84	0.17	0.1	0.55	-4.2	0.57	-4.1	1.41	25
677	132	5.1	4.85	0.17	0.1	1.15	1.1	1.24	1.8	0.75	28
693	144	4.8	4.85	0.16	0.1	0.56	-4.1	0.58	-4.1	1.49	23
696	144	4.8	4.88	0.14	0.1	0.51	-4.8	0.57	-4.2	1.46	20
710	144	4.9	5.02	0.01	0.1	0.95	-0.3	1.02	0.2	1.02	27
716	144	5	5.08	-0.05	0.1	0.98	-0.1	0.87	-1.1	1.06	14
719	144	5	5.11	-0.07	0.1	0.62	-3.5	0.62	-3.6	1.46	7
721	144	5	5.13	-0.09	0.1	0.9	-0.7	1.18	1.4	0.79	2
722	144	5	5.14	-0.1	0.1	0.7	-2.6	0.77	-2	1.23	1
722	144	5	5.14	-0.1	0.1	0.61	-3.5	0.66	-3.1	1.32	15
728	144	5.1	5.2	-0.16	0.1	1.21	1.5	1.06	0.4	0.99	4
730	144	5.1	5.22	-0.18	0.1	1.09	0.7	1	0	1.01	16
737	144	5.1	5.28	-0.24	0.1	0.94	-0.4	0.83	-1.4	1.19	8
739	144	5.1	5.3	-0.26	0.1	1.1	0.8	1.01	0	0.94	5
743	144	5.2	5.34	-0.3	0.1	0.7	-2.5	0.68	-2.9	1.27	13
765	144	5.3	5.56	-0.5	0.1	0.68	-2.7	0.7	-2.6	1.24	12
499	84	5.9	6.18	-1.13	0.15	1.05	0.3	0.82	-0.9	1.15	26
Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Raters
681.6	138	5	5.02	0	0.1	0.86	-1.3	0.85	-1.4		Mean
87.6	18.3	0.3	0.38	0.36	0.01	0.22	1.9	0.2	1.8		S.D. (Populn)

Rater separation index: 40.90; separation reliability: 0.92; fixed chi-square p = 0.00

performances

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Performances
1214	138	8.8	8.8	7.53	0.21	0.91	-0.5	0.88	-0.7	1.07	5 RINA
1184	138	8.6	8.58	6.54	0.16	0.87	-1.2	0.86	-1.3	1.19	6 ESTERE
1065	138	7.7	7.72	4.24	0.12	0.83	-1.3	0.78	-1.6	1.19	10 ELLI
1079	144	7.5	7.5	3.85	0.11	0.91	-0.6	0.91	-0.7	1.09	14 KARINE
1053	144	7.3	7.32	3.57	0.1	1.15	1.2	1.13	1.1	0.87	13 KIM
982	138	7.1	7.15	3.33	0.1	0.69	-2.9	0.66	-3.3	1.31	18 RAQUEL
949	138	6.9	6.87	2.97	0.09	1.14	1.2	1.12	1	0.82	9 WILMA
920	138	6.7	6.7	2.77	0.09	0.75	-2.2	0.74	-2.4	1.29	17 AMALIA
922	138	6.7	6.67	2.74	0.09	0.96	-0.3	0.94	-0.4	0.95	1 MARTA
818	138	5.9	5.97	1.98	0.09	0.69	-2.8	0.7	-2.7	1.34	21 DIMITRIO
816	138	5.9	5.9	1.91	0.09	0.66	-3.2	0.67	-3.1	1.36	7 AGATA
695	138	5	5.03	1.07	0.08	0.66	-3.1	0.68	-3	1.31	8 SIMON
682	138	4.9	5	1.04	0.08	0.57	-4.2	0.58	-4.1	1.44	22 VERONICA
604	138	4.4	4.38	0.48	0.08	0.7	-2.7	0.7	-2.7	1.33	2 MEGUMI
524	138	3.8	3.86	0.04	0.08	0.98	-0.1	0.99	0	1.07	19 STEFANIE
492	138	3.6	3.62	-0.16	0.08	0.76	-2.3	0.77	-2.1	1.23	20 EWA
469	138	3.4	3.45	-0.31	0.08	1.44	3.6	1.46	3.7	0.46	24 DESIREE
439	138	3.2	3.22	-0.51	0.08	1.17	1.5	1.16	1.3	0.91	23 CRAIG
402	138	2.9	2.94	-0.77	0.09	0.64	-3.6	0.7	-2.9	1.16	15 SOPHIE
286	138	2.1	2.1	-1.87	0.11	0.49	-4.3	0.49	-4.3	1.37	16 MOHAMMAD
243	138	1.8	1.76	-2.59	0.13	0.91	-0.5	0.88	-0.8	1.05	12 MALEGIO
184	132	1.4	1.4	-3.66	0.16	0.95	-0.3	0.93	-0.5	1.06	4 ALLISON
168	132	1.3	1.28	-4.14	0.18	0.99	0	0.93	-0.4	1.08	3 DIANA
168	138	1.2	1.23	-4.4	0.19	0.83	-1.1	0.79	-1.4	1.15	11 BRUCE
Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Performances
681	6 138	.0 4.9	4.93	1.07	0.11	0.86	-1.3	0.85	-1.3		Mean
333	2 2	.4 2.4	2.37	3.05	0.04	0.22	1.9	0.21	1.9		S.D. (Populn)

Performance separation index: 34.68; separation reliability: 1; fixed chi-square p = 0.00

Appendix IV: rater and respondent groups used in the analysis

rater characteristics

4 gender
5 native speaker
6 curriculum planner
7 *CEFR* expert
8 *CEFR* seminar participant
9 CELI oral examiner
10 CELI item writer
11 have *CEFR* knowledge*
12 have CELI knowledge*

respondent characteristics

13 gender
14 country of origin
15 European*
16 speak European language*

*these groups were constructed by merging or simplifying other groups in order to ensure more data points within each group