All different – all equal?

Towards cross-language benchmarking using samples of oral production in French, German and Italian

Gilles Breton, Giuliana Grego Bolli, Michaela Perlmann-Balme

Abstracts

The aim of this paper is to report about how to rate samples of oral performances according to the levels of the *Common European Framework of Reference for Languages* (CEFR). The paper summarizes the outcomes of three benchmarking seminars. The French samples are provided by Centre International d'Etudes Pédagogiques (CIEP), the German samples by the Goethe-Institut and the Italian samples by CVCL of the Università per Stranieri di Perugia. The method applied was derived from the Council of Europe's document *Manual for Relating Language Examinations to the Common Framework of Reference for Languages* released in 2003. The paper outlines how to carry out such events, and which problems to be aware of.

Ziel dieses Beitrags ist aufzuzeigen, wie mündliche Leistungen in die Niveaus des *Gemeinsamen europäischen Referenzrahmens* für *Sprachen* (GER) eingestuft werden können. Er fasst die Ergebnisse dreier Seminare zum „benchmarking" von mündlichen Kandidatenbeispielen zusammen. Die französischen Beispiele dazu stammen vom Centre International d'Études Pédagogiques (CIEP), die deutschsprachigen Beispiele vom Goethe-Institut und die italienischen von CVCL der Università per Stranieri di Perugia. Methodische Grundlage ist das vom Europarat 2003 veröffentlichte *Manual for Relating Language Examinations to the Common Framework of Reference for Languages*. Der Beitrag beschreibt, wie eine solche „Benchmarking Konferenz" durchgeführt werden kann und welche Probleme sich ergeben.

Cette contribution a pour principal objectif de montrer comment des exemples de production orale peuvent être situés par rapport aux niveaux du *Cadre Européen Commun de Référence pour les Langues* (CECR). Elle résume les résultats d'un séminaire de calibrage d'examens oraux. Les exemples français proviennent du Centre International d'Études Pédagogiques (CIEP), les exemples allemands du Goethe-Institut et les exemples italiens du CVCL de l' Università per Stranieri di Perugia. Le *Manuel pour relier les examens de langues au CECR,* publié par le Conseil de l'Europe en 2003, en est la base méthodologique. La contribution décrit comment organiser de tels séminaires afin d'aider d'autres organisations à faire de même dans l'avenir. Elle montre enfin comment d'autres organisations peuvent mettre en place ce type de séminaire de calibrage.

Dr. Michaela Perlmann-Balme
Goethe-Institut
Dachauer Str. 122
80637 München
Tel. +49 89 15921-382
Fax: +49 89 15921-102
E-Mail: perlmann-balme@goethe.de

Giuliana Grego Bolli
Prof. Associato di Linguistica Applicata
Direttore del CVCL (Centro per la Valutazione e le Certificazioni Linguistiche)
Palazzina Lupattelli
Via XIV Settembre 75
06124 Perugia
Tel. +39 0755746719
E-Mail: giuliana.bolli@gmail.com

Gilles Breton
1, avenue Léon Journault
92318 Sèvres CEDEX
Telephone: 01 45 07 60 84 - Fax: 01 45 07 60 01
E-Mail: crd@ciep.fr

Introduction

Most people who have learned more than one foreign language at school or in an educational institution have a feeling: There are strong traditions in marking or rating language proficiency. What is a "bad mistake" for teachers of French might not be penalized as harshly by teachers of English or German. In the light of the growing importance of foreign language learning in Europe it is important to provide orientation how to rate performances in a foreign language in such a way that language specific traditions do not play the main role. What is needed is a common basis. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* provides the theoretical basis. The next step is a need for illustrative samples of what this basis means in the practical application across the different European languages.

This paper can be seen as the first step towards a cross language benchmarking. By using the procedures provided by the preliminary pilot version of the manual published by the Council of Europe in 2003 – *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* – language experts rated a set of video samples of three widely spoken languages in Europe – French, German and Italian – in a series of conferences. They analysed the dimensions of spoken language which were defined by the *CEFR* and applied the scales provided there to these video samples. These language samples represent each level of the *CEFR*. They can be used by other language experts across Europe and the world in the process of comparing their own performance samples. These samples set standards and as such can be used to compare "local samples" such as placement or achievement tests of schools and universities.

These benchmarked performances in French, German and Italian are available for experts in the education sector. They can be used in training of teachers and examiners because they allow a practical approach to the understanding of what a level of proficiency means. Secondly, they can be used to standardise rating in educational institutions where language teachers are involved in rating oral proficiency in foreign languages. While at present teacher often use their own concept of a proficiency level and therefore differ from one teacher to the other, the benchmarking samples can be used to standardise these ratings and thus make them fairer and more reliable.

The procedures of conducting these conferences were exactly the same in each of the three languages. These first benchmarking events were conducted by three established examination providers and members of the *Association of Language Testers in Europe* (*ALTE*): For French it was conducted by the Centre International d'Etudes Pédagogiques (CIEP), for German by the Goethe-Institut [1], for Italian by CVCL of the 'Università per Stranieri di Perugia'. All three conferences took place in 2005. Meanwhile there was a fourth conference conducted in Portuguese in the same way in 2006 as well as a conference regarding samples of young learners in English, French, German, Italian and Spanish in 2008. This paper describes not only the main results of the 2005 conferences but also how to carry out such events in order to help organisations in other languages as well as other institutions conducting such events in order to link oral performances to the CEFR.

The methodology of the Manual as a basis for the organisation of benchmarking events

The production and the publication of the CEFR was just the first step in the implementation of a common framework in the field of language learning, teaching and assessing. North and Schneider (1998: 243) emphasised the importance of ensuring a common interpretation of the scale levels through standardised samples of performances.

1 For German there was a second conference organised by Österreichisches Sprachdiplom (ÖSD) in Vienna in December 2006. Benchmarked samples of this event are published alongside with samples provided by the Goethe-Institut on the DVD *Mündlich* (2008).

The very general brief of the draft manual: *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe 2003) was to support the process of standardising, the interpretation of the levels, providing guidelines and suggesting procedures. These procedures are part of a more complex methodology for relating language examinations to the CEFR, providing evidence of the claim made by the majority of examination providers of the linkage of the examinations to the CEFR levels. The above mentioned procedures are presented in four sets in the *Manual* (Council of Europe 2003):

- Familiarisation (Chapter 3) with the CEFR in order to ensure an in-depth knowledge of the levels descriptors using general and specific tables.
- Specification (Chapter 4) of the coverage of the examination in relation to the CEFR categories and descriptors at the different levels.
- Standardisation (Chapter 5) of the interpretation of the CEFR levels through familiarisation activities and a specific training with examples performances calibrated to the CEFR scale using common criteria (CEFR Table 3) followed by benchmarking local performances examples to the framework levels. In order to do that DVDs of performances illustrating the CEFR levels have been produced in French, German and Italian.
- Empirical validation (Chapter 6) of the results of an examination through a technical corroboration of the linking claimed through specification and standardisation. In order to achieve this, a correlation to an external criterion should be demonstrated. This criterion can either be a test already calibrated to the CEFR or judgements (by teachers, examiners, language experts) based on the CEFR descriptors.

The standardisation of judgments, as described in Chapter 5 of the *Manual*, has been used as common basis in organising the benchmarking events run so far in France, Germany and Italy in order to reach a common interpretation of the CEFR level descriptors in relation to the spoken performance in three different languages. As a first result of these first national benchmarking events, three DVDs have been produced offering the first examples of spoken performances in French, German and Italian. These performances have been linked, by experts' judgment, to the CEFR levels during the above mentioned benchmarking seminars.

The seminars organised so far were intended also to demonstrate how it is possible to use the CEFR scale descriptors and work with them concretely (Grego Bolli 2008), as well as to encourage other institutions to organise local benchmarking events in order to link local examples of spoken performances to the already calibrated samples.

Organising and conducting a benchmarking seminar

A benchmarking seminar addresses both the horizontal and the vertical dimension of the *Framework*. In fact, it implies a cross comparison about performances compared to the vertical dimension, which means levels, defined by descriptors. The horizontal dimension rates qualitative descriptors of language performance.

Organising and conducting a benchmarking seminar requires one to systematically apply a set of procedures taken from Chapter 5 of the *Manual*. These procedures, applied during the seminars run in Sèvres, Munich and Perugia, can be basically divided into three different stages: before, during and after the seminar.

Before the seminar, samples of spoken performances are collected according to a specific format. In order to do this, tasks[2] at different levels are selected and students at different levels of language proficiency are recruited. A number (around 25) of experts (teachers, examiners, etc), representative of different contexts, are invited to participate and the logistics of each national event is carefully planned.

During the event, the coordinator or the person responsible for each event provides the participants with all the materials to be used during the seminar according to what is indicated in Chapter 5 of the *Manual*. In particular: familiarisation activities, as described in Chapter 5, are run, rating instruments are distributed and voting procedures are set up.

The main rating instruments used during the three above mentioned seminars were: Table 5.4, translated into French, German and Italian from the *Manual* and Table 3 from the national versions of the CEFR.

Participants were also furnished with a supplementary grid, translated into the three languages, giving descriptors for the "plus" levels: A2+, B1+ and B2+ already mentioned. The addition of these three "plus" levels resulted in 9 levels overall. "These 9 levels reflect the linear scale produced in the Swiss research project" (North & Lepage 2005: 5). The participants were encouraged to consult other relevant scales before rating and voting: Overall Spoken Production, Overall Spoken Interaction, Sustained Monologue: Describing experience. The rating material was distributed to all the participants after the Familiarisation activities.

Rating was first done on paper adapting Form B2 (analytic rating form) from the *Manua*l in two versions with two different colours in order to distinguish the votes cast before and after discussion. Votes were then recorded electronically with CEFR levels corresponding to buttons on the keypad provided by a specialised French agency. According to the conclusions drawn on the report of the first seminar in Sévrès (North & Lepage 2005), it was decided, both in Munich and Perugia, to reverse the original procedure, i.e. going from a global assessment (using Table 5.4) to an analytic one (using Table 3 and the "plus" levels grid).

Before participants started to watch the sequences some initial training was done by watching three sequences intended to represent the three broad CEFR levels: B, C, A (in the order they were presented). Even if only the first evaluation sequence was really used as training material, more time was spent on each of the three finding out according to which descriptors the participants assigned the level, then commenting on the analysis of the Excel table shown after the first voting, reading descriptors and discussing the difficulties and doubts encountered by the participants. The intention was to use the first three sequences as 'broad' points of reference for the following rating. As North and Lepage pointed out, this task aimed "to get participants accustomed to the rating instruments and to the process of first recording judgments on paper and then voting electronically, but it also served to show the degree of agreement in the group" (North & Lepage 2005: 7).

Participants were asked to watch a few minutes of the production phase of the first two learners and decide first, as a group, on the broad level (A, B, C) of each learner. During the following discussion participants were asked, by the seminar coordinators, to read the descriptors they focused on in order to link the performances to the level both in relation to the scales and to the criteria in the grids. The descriptors were read, commented on and discussed in relation to each performance; this was done in the context of the descriptors for the level above and below. The training phase was quite successful in clarifying both procedures and criteria as much as possible.

2 The tasks used in all the three events were a further development of those developed for standardisation videos in the Swiss research project that had produced the CEFR levels and descriptors. These tasks had been used for English video circulating in April 2004 and were recommended in the Council of Europe's Brief of Recording."(North, B. & Lepage, S. 2005. Seminar to calibrate examples of spoken performances in lane with the scales of the *Common European Framework of Reference for Languages,* Strasbourg: Council of Europe, 5).

The same basic pattern of rating was then followed in the three benchmarking events:

1. Watching the sequence (Production: Learner A; Production: Learner B; Interaction between Learner A and B).
2. Consulting Table 5.4 (and other scales provided in addition), reflecting and rating the Global Level for both learners individually on Form B2.
3. Consulting criteria grids, reflecting and recording on Form B2 the rating for both learners for the 5 criteria of the grids: Range, Accuracy, Fluency, Interaction and Coherence.
4. Electronic voting: individual votes: Range, Accuracy, Fluency, Interaction and Coherence, Global Level (Livello Globale).
5. Viewing histogram of the individual global judgement.
6. Starting the discussion in small groups.
7. Reporting the results in plenary discussion
8. Viewing the Excel table of the individual votes for each of the five criteria for learner A. Plenary discussion. View of the Excel table of the individual votes for each of the five criteria for learner B. Plenary discussion[3].
9. Reflecting and rating individually the final Global Level for learner A and for Learner B on Form B2 (version with a different colour from the one used for the voting before discussion).
10. Electronic voting: Global.
11. Viewing the histogram of global judgements after the discussion.

Only the most problematic recordings were shown twice.

During the seminar in Perugia, it was decided not to split into small groups for the discussion, principally because of time constraints.

After the events, DVDs in French, German[4] and Italian were edited by each of the institutions responsible for the organisation of the seminars, providing samples of oral performances linked by experts' judgement to the CEFR levels. The results of the ratings have been statistically analysed by Cambridge ESOL and the results commented on and published on the Council of Europe website.

The samples of oral proficiency

The procedure of selecting the samples shown during the presentation of the cross language benchmarking experiences as well as the format and the tasks chosen give a good opportunity of reflecting upon the notion of difference and equality when assessing different languages.

The first step was the selection of the samples to be used during the seminars.

For the CIEP seminar at Sèvres in December 2004[5], the learners filmed were selected in a systematic way: teachers' evaluations, questionnaires and test results (using the new CEFR based DELF DALF).The performances were viewed by an expert group who rated them onto the CEFR levels and discussed each sample before making a final selection for the seminar.

For the Goethe Institute's seminar in Munich, October 2005[6], as well as for the CVCL seminar at the Università per Stranieri di Perugia in December 2005[7], a similar procedure was followed: class observations, teachers' evaluation, learners interviewed by the

3 For most of the sequences.

4 The German DVD contains benchmarked samples from two conferences: Munich 2005 and Vienna 2006. They are published by Langenscheidt *Mündlich* (2008).

5 Seminar to calibrate examples of spoken performances in line with the scales of the CEFRL CIEP, Sèvres, 2-4 December 2004. Report by Brian North and Sylvie Lepage.

6 Seminar to calibrate samples of spoken performances to the CEFRL, Goethe-Institut, Munich, 19th-22nd October 2005. Report by Sibylle Bolton. For the Vienna conference: Benchmarking Conference German: Spoken Performance, Vienna, 7-10 December 2006, organised by the ÖSD, Report on analysis of rating data by Guenther Sigott.

7 Seminar to calibrate examples of spoken performances in Italian L2 to the scales of the CEFRL Università per Stranieri di Perugia; 16th-17th December 2005. Report by Giuliana Grego Bolli.

seminar's coordinators to obtain a clearer idea of the possible performances in relation to the CEFR levels.

As to the tasks, their origin and the format, these were also similar for French and Italian with slight differences. The tasks did not follow the same pattern in German.

For French and Italian, the tasks filmed were a further development of those developed for standardisation videos in the Swiss research project that had produced the CEFR levels and descriptors. They had been used for the English video circulated in April 2004 and were recommended in the Council of Europe's "Brief of recording".

Each recording shows two learners, with no native speaker examiner/interlocutor for French, but one for Italian being present not as an interlocutor but to introduce the topics and help in case of serious breakdown in the communication. The Goethe-Institut decided to introduce a native-speaker interlocutor to probe the learners should this become necessary.

These recordings consist of three phases:

- a production phase by the first learner delivering a sustained monologue, which may be followed by questions from the other learner; in the German case, the production phase may be followed by questions from the native-speaker interlocutor.
- a similar production phase by the other learner;
- a phase of spontaneous interaction between the two learners.

The same test format was used at all levels. The production phase is semi-prepared, i.e. the learners can choose a topic and reflect for 5 to 10 minutes on what they want to say. The interaction phase is spontaneous, elicited by cards containing discussion questions with an element of learner choice for the topics. The two learners can discard topics that do not interest them. The format is more standardised in German and the topic is given at C1 and C2 levels.

For French, recordings with a total duration of 12 minutes were shown in their entirety whereas longer recordings were shortened to extracts of 3-4 minutes for each of the three phases. All the Italian recordings were shown during the seminar (average length 15 minutes).

The elements given above show to which extent the procedures of selection of the learners and the tasks were similar as far as there is a systematic reference to the CEFR levels. The differences mainly concern the presence or absence of an interlocutor as well as the opportunity for the learners to choose the topics. The performances are in fact considerably influenced by the format used.

The cross language benchmarking seminar that was organised at the CIEP in Sèvres in June 2008 took into account the experience of the three previous seminars using the same procedures for the performances and the samples in English, French, German, Italian and Spanish: no interlocutor was used except in case of serious breakdown of communication (which occurred during some filming, because the learners were teenagers); the same format was used for all levels, including the opportunity of choosing the topic, which is of great importance.

Empirical results

The aim of this part is to summarize the main results of the three Council of Europe benchmarking seminars, organised by the members of ALTE: Sèvres (2005), Munich (2005) and Perugia (2005). The following results are extracted from the reports of Neil Jones on the conferences in Sèvres, Munich and of Michael Corrigan about the conference in Perugia.[8]

Jones (2006, 2007) and Corrigan (2006) conducted analyses on data from each seminar and a selection of charts and tables of this material are given here.

Particularly interesting are those results which hint at issues which are related to more than just one language or just one particular circumstance. There are three issues emerging which are of general importance:

- The level of agreement of raters at each CEFR level
- The severity or leniency in the application of individual rating criteria
- The feasibility of using the 9 levels of the CEFR

The level of agreement of raters at each CEFR level

During all of the conferences there were two judgements collected from all the raters. First, there was an individual and independent rating. The result of this first voting was collected then followed by a discussion. The aim of this discussion was to find a consensus about the level of the oral performances which were shown. After this discussion all raters made a second judgement. The results of the first and second or final votes were:

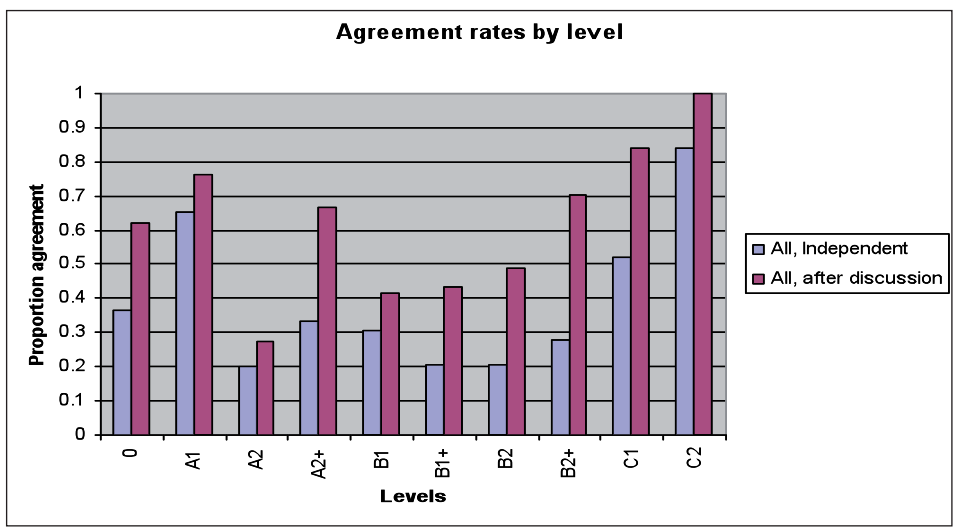

Figure 1: Sèvres - Agreement rates by level before and after plenary discussion

8 Jones, Neil (2006): Seminar to calibrate examples of spoken performance, Goethe-Institut, Munich, November 2005, Report on analysis of rating data, Draft version, 18 September 2006.

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

Corrigan, Michael (2007): Seminar to calibrate examples of spoken performance, Università per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica), Perugia, 17-18 December 2005. Report on the analysis of the rating data.

Both reports can be downloaded from the web page of the Lanugage Policy Division, Council of Europe: http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html

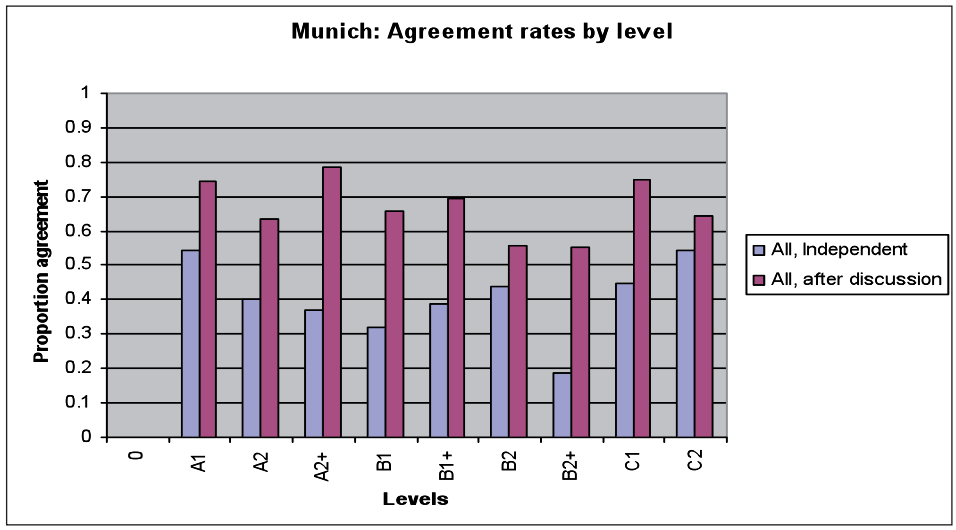

Figure 2: Munich - Agreement rates by level before and after plenary discussion

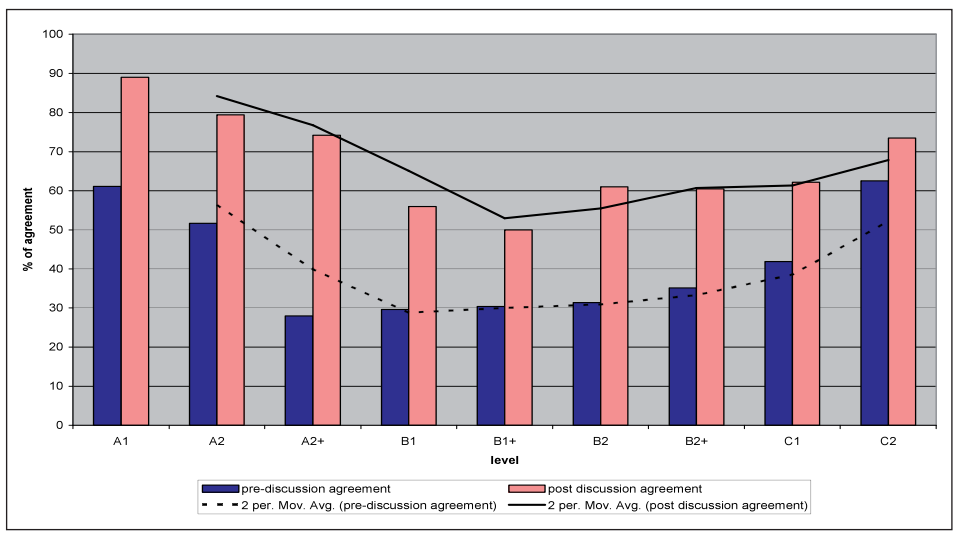

Figure 3: Perugia - Agreement rates by level before and after plenary discussion[9]

The graphs show that the post discussion levels of agreement were higher at all three conferences. Since some participants were less familiarized with the use of the CEFR than others this is not surprising. This has two consequences. First, it is important to devote enough time to the familiarization during the benchmarking event. Secondly, it is therefore important to make sure that the experts invited to these benchmarking events have as much experience which such votings as possible. Those participants who had experience with conducting oral examinations and with using the CEFR could give more convincing arguments for their votes. The process of reasoning and supporting the argument by giving examples influenced the raters in their interpretation of the levels.

The diagrams also show that agreement at either end of the scale is easier to achieve because there are fewer alternatives. However, it should be recognised that agreement on the levels in the middle – B1, B1+, B2 and B2+ is still high. It only appears low in relation to those levels at either end of the scale.

9 We would like to thank the authors of the reports on the benchmarking events in Sèvres and Munich, Neil Jones, and of Perugia, Michael Corrigan, for the first analysis of the data. Both have provided the graphs which are presented here. Final touches to the diagrams in this paper were done by Michael Corrigan.

The severity or leniency in the application of individual rating criteria

A second interesting result of the conferences was the different behaviour of the raters applying the rating criteria of table 3 of the CEFR: "Qualitative aspects of spoken language use". The five criteria are: range, accuracy, fluency, interaction and coherence. It is possible to see how lenient or severe raters were for each criterion. Shorter bars indicate greater leniency, longer ones greater severity. 0 marks the centre of the scales. The negative numbers show more leniency.

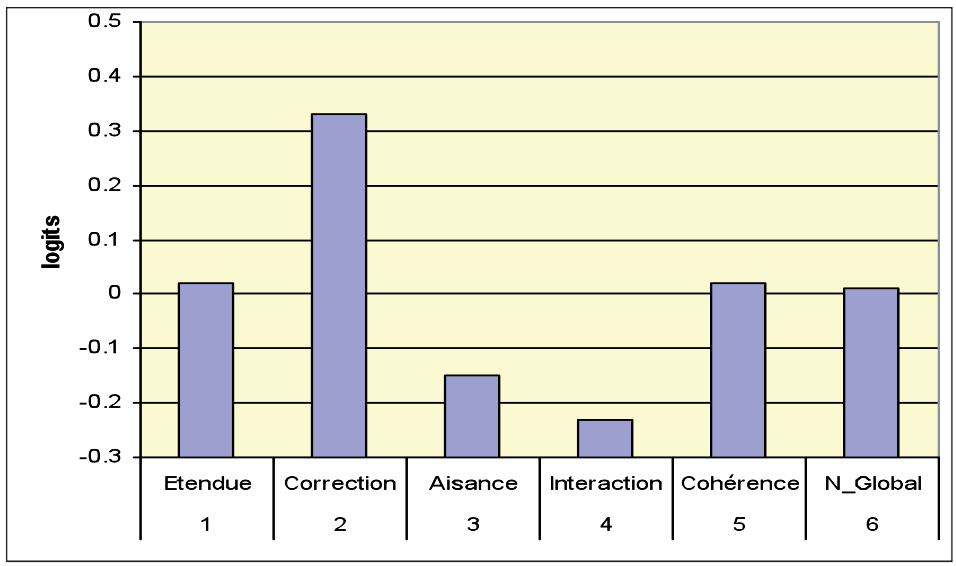

Figure 4: Sèvres - Relative difficulty of rating criteria (FACETS analysis[10])

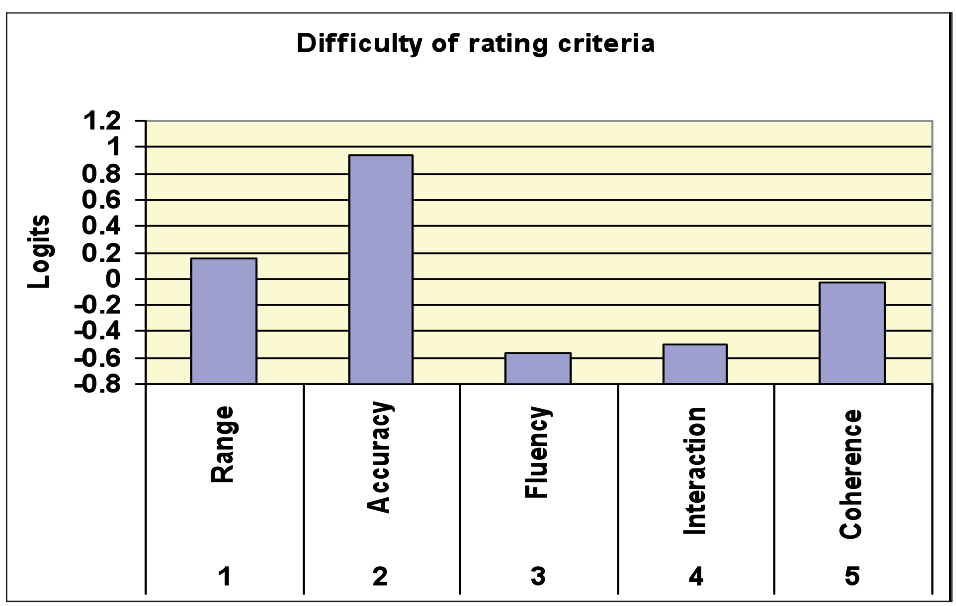

Figure 5: Munich - Relative difficulty of rating criteria (FACETS analysis)

10 FACETS is a Rasch model statistical computer program published by John M. Linacre 2005.

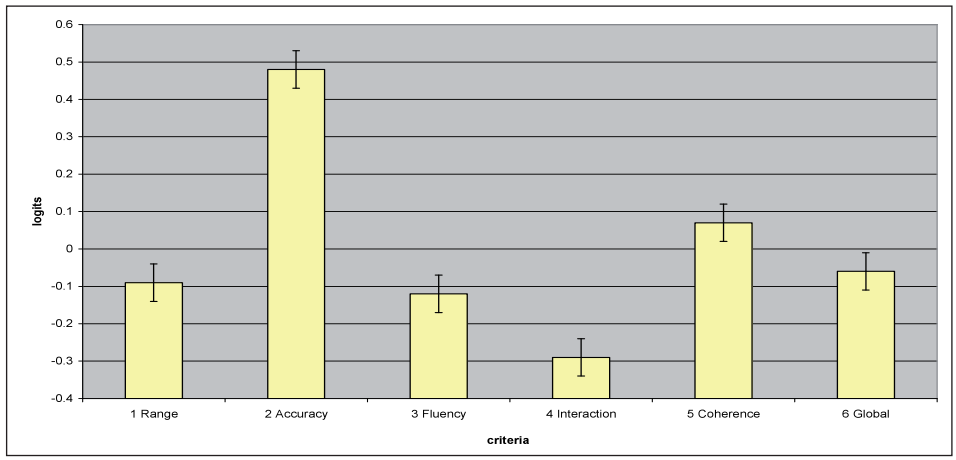

Figure 6: Perugia - Relative difficulty of rating criteria (FACETS analysis)

The criterion "accuracy" was dealt with in the most severe way by raters in all three languages. That means the raters chose a lower level for accuracy than for the other criteria. In contrast to that the criterion "interaction" was rated most leniently by raters who voted on performances in French and Italian, in German this criterion became the second most lenient. "Fluency" was also rated rather leniently. Raters considered the descriptors for higher levels as appropriate.

This behaviour of raters shows that the tradition of measuring knowledge of grammatical rules is still strong among raters for all three languages. According to this tradition counting mistakes is still alive among language teachers and experts. Accuracy in the sense of correctness weighs still more for some raters than the "soft" criteria of interaction and fluency do.

The difficulties of raters using the 9 levels of the CEFR

When conducting the benchmarking conferences the use of the so-called plus-level was focussed on more than was expected. In a number of cases the raters felt that it was not enough to use the six "full" levels: A1, A2, B1, B2, C1 and C2. In order to differentiate the performances in yet finer shades, the three plus-levels A2+, B1+ and B2+ were applied. Moreover, in Sevres and Perugia, there was a level below A1.

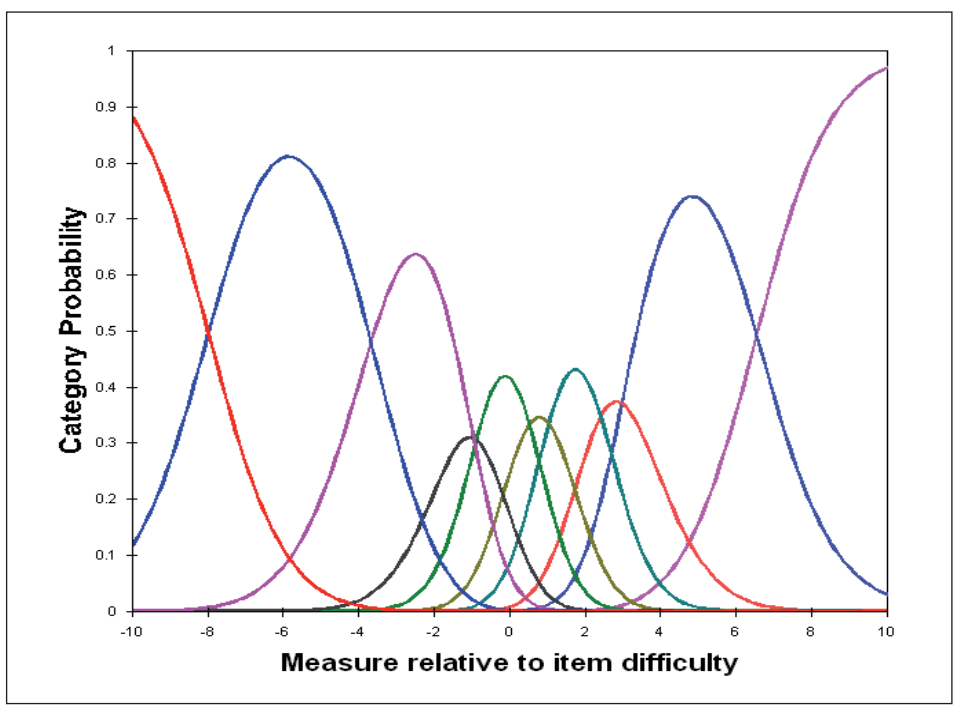

Figure 7: Sèvres – 10 point scale

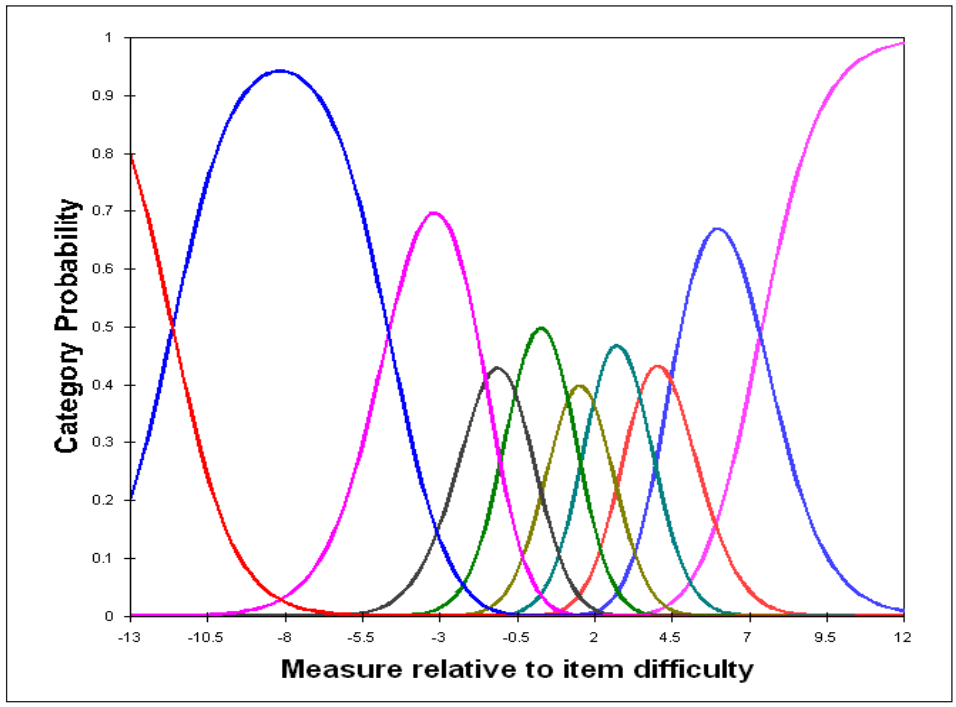

Figure 8: Perugia – 10 point scale

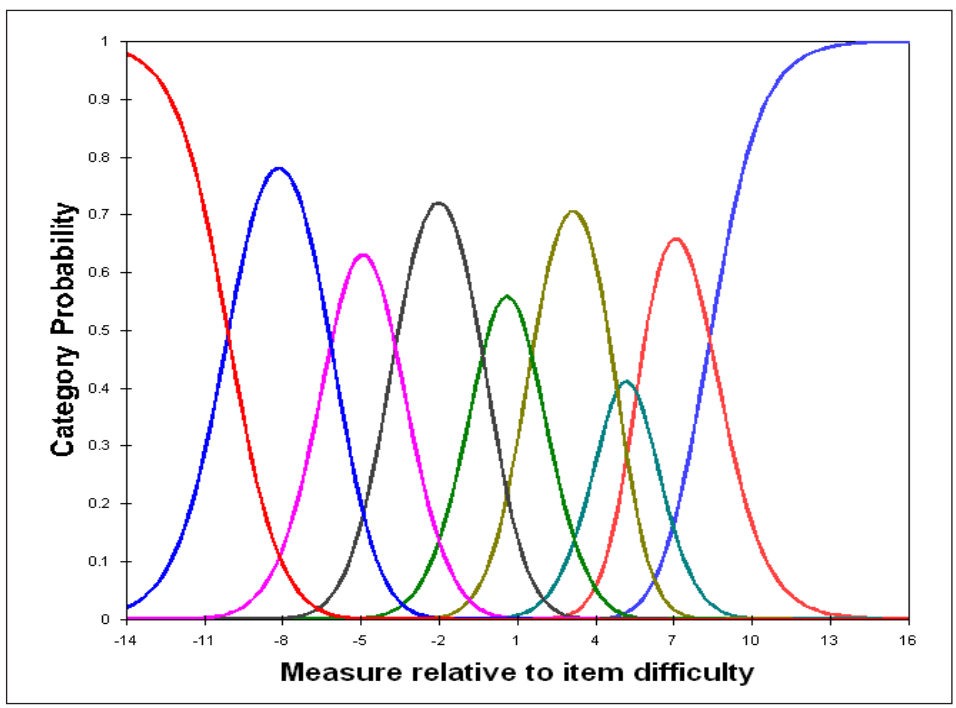

Figure 9: Munich – 9 point scale

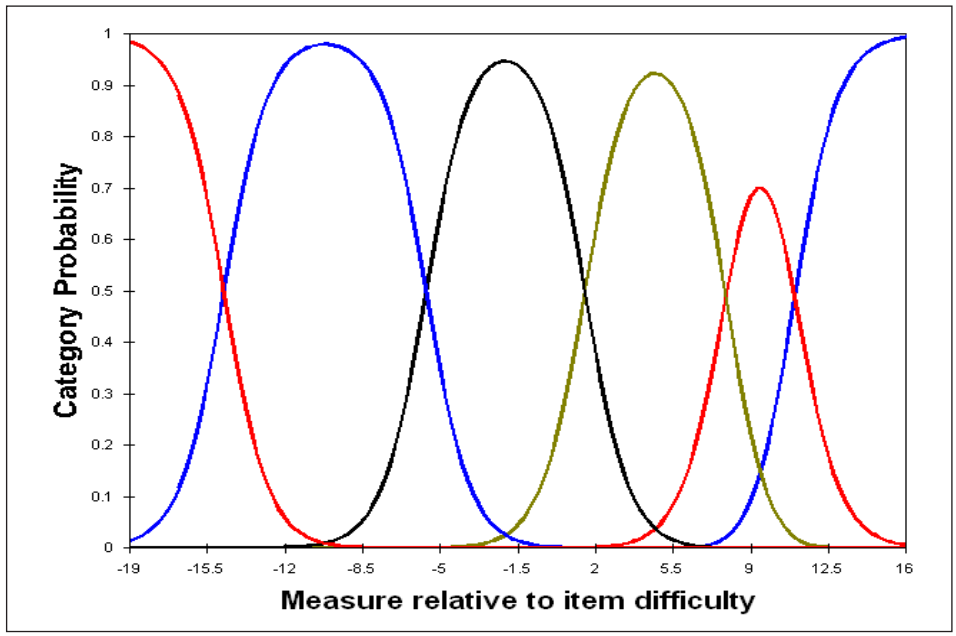

Figure 10: Munich – 6 point scale

The colours for the diagrams are as follows:

Sevres and Perugia		Munich (both scales)	
1 A0	red	A1	red
2 A1	blue	A2	blue
3 A2	pink	A2+	pink
4 A2+	black	B1	black
5 B1	green	B1+	green
6 B1+	olive	B2	olive
7 B2	turquoise	B2+	turquoise
8 B2+	red ii	C1	red ii
9 C1	blue ii	C2	blue ii
10 C2	pink ii		

It is important to obtain clear agreement between experts about which rating should be given for which level of performance. When using the Many Facet Rasch Measurement (MFRM)[11] model clear agreement about the meaning of each level will translate into a clearly-delineated scale. MFRM scales for each event are shown in figures 7, 8 and 9 and a key is provided after figure 10. In each case, for each rating category (A1, A2, A2+, B1, B1+, B2, B2+, C1, C2), the ability level needed to get a particular rating (horizontal axis) is plotted against the likelihood of obtaining it (vertical axis). On the horizontal axis, ability is lowest on the left and on the vertical axis, likelihood is highest at the top of the diagram. The most likely rating for the lowest level of performance can be seen as A1. For the next lowest ability range, as would be expected, A2 is the

11 MFRM can measure candidate/performance ability, task difficulty and rater severity among other aspects of such rating exercises. It is based on a mathematical model which uses the concepts of difficulty and ability to describe the likelihood, in the example of the current research, of each candidate's performance being given a particular rating. In the model, difficulty is broken down into *facets*, such as the severity of the rater and the difficulty of the task. A single scale is used to place (or measure) each element of each facet (e.g. each rater in the rater facet, each candidate/performance in the candidate/performance facet) on a single ability/difficulty scale. When conducting MFRM, there are certain expectations about the way the rating scale is modelled using the ability/difficulty scale. Firstly, each rating should be, for some level of ability, the most likely rating. Secondly, ratings should be monotonic, or the most likely rating *in order*, so lower ratings should be more likely at lower levels of ability and higher ratings at higher level. Greater agreement when using rating scales should lead to a more clearly-defined, monotonic scale when it is modelled using MFRM.

most likely rating. The other ratings follow sequentially across the entire ability range, usually peaking as the most likely rating for a given ability range.

Problems are, however, evident in Figures 7 to 9. For figures 7 and 8, rating categories B1 to C1 are compressed and overlap each other more. In figure 7, the B1 category is never the most likely. In Figure 9, the B1+ and B2+ categories are compressed. These features indicate that the raters found it far more difficult to agree clearly on the rating of these levels. That these problems occurred in the region of the scale containing the plus levels is not surprising. Distinguishing finer shades of proficiency is always more difficult than defining broader categories. A further analysis was done using the same data represented in figure 9 with fewer categories. The number of categories were reduced by collapsing the plus categories into the main categories (i.e. those ratings in A2+ were included in A2, those in B1+ were included in B1, those in B2+ were included in B2). Figure 10 shows that a more clearly-defined system of levels results.

Experience from the Munich conference showed that when it came to the voting the plus-levels functioned more and more as a compromise level for performances that were felt slightly above or below the level.

Other interesting results which came out of the reports of Jones (2005, 2006) and Corrigan (2007) concerned the raters' behaviour: Statistics suggest that the raters were learning or becoming more familiar with the rating criteria, the format of the event and of the examples of oral production as the conference continued. It also reveals that different groups of raters – such as language experts versus non language experts – did not differ in their voting. There is a much greater difference between the way all groups voted on criteria than on one another. This suggests that the influence of groups on voting was not as important as the influence of the different criteria.

Further steps

The events in Sèvres, Munich and Perugia, which were described here were followed by a cross languages benchmarking event in June 2008 organised by the CIEP in Sèvres. Building on the experience of the language specific conferences of 2005 there were several changes: Firstly, English and Spanish were integrated so that the DVD published by the Council of Europe (2009) contains samples in five European languages, secondly, the language samples were taken from young learners aged 13 to 18 years old, thirdly, the about fifty experts from the Council of Europe, Goethe Institute, Cambridge ESOL, CVCL of the 'Università per Stranieri di Perugia' and Instituto Cervantes rated samples in two languages, so that there was a link between the ratings across languages. There is also a need for some cross language benchmarking of written production which can be seen as the next most important step in this process of standardising the rating of foreign language performances across Europe.

Bibliography

Bolton, S. (2006). *Seminar to calibrate samples of spoken performances to the Common European Framework of Reference for Languages*. http://www.coe.int/T/DG4/Portfolio/main_pages/Report%20Seminar%20in%20German.pdf (accessed on 08.07.2010).

Bolton, S., Glaboniat., M., Lorenz, H., Perlmann-Balme, M., Steiner, S. (2008) *Mündlich. Mündliche Produktion und Interaktion Deutsch. Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens*. Berlin, München: Langenscheidt.

Commission of the European Communities (2007). *Final Report, High Level Group on Multilingualism*. Luxembourg: Office for Official Publications of the European Communities. http://ec.europa.eu/education/policies/lang/doc/multireport_en.pdf (accessed on 08.07.2010).

Corrigan, M. (2007). *Seminar to calibrate examples of spoken performance. Università per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica. Report on the analysis of the rating data*. http://www.coe.int/T/DG4/Portfolio/documents/Report_Seminar_Perugia05.pdf (accessed on 08.07.2010).

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Manual Preliminary Pilot Version. DGIV/EDU/LANG (2003) 5, Strasbourg: Council of Europe.

Council of Europe (2009). Mündliche Leistungen: Beispiele für die 6 Niveaustufen. Paris: CIEP. www.ciep.fr (DVD).

Franceschini, R. & Miecznikowski, J. (eds.) (2004). *Leben mit mehreren Sprachen. Vivre avec plusieurs langues. Sprachbiographien / Biographies langagières*. Bern: Lang.

Grego Bolli, G. (2008). Progetti europei: nuove prospettive sulla scia del Quadro Comune Europeo di Riferimento. In Ciliberti, A. (a cura di) *Un mondo di italiano*. Perugia: Guerra Edizioni, 25-48.

Jones, N. (2005). *Seminar to calibrate examples of spoken performance: CIEP Sèvres, 02-04 December 2004. Report on analysis of rating data*. http://www.coe.int/T/DG4/Portfolio/documents/SevresreportNJ.pdf (accessed on 08.07.2010).

Jones, N. (2006). *Seminar to calibrate examples of spoken performance, Goethe-Institut, Munich, November 2005, Report on analysis of rating data*, Draft version, 18 September 2006. (unpublished ms).

Lepage, S. & North, B. (2005). *Exemples de productions orales illustrant, pour le français, les niveaux du Cadre européen commun de réference pour les languages*. Paris: CIEP.

North, B. & Schneider, G. (1998). Scaling Descriptors for Language Proficiency Scales. In: *Language Testing* 15, 217-262.

North, B. & Hughes, G. (2003). CEF Performance Samples. English (Swiss Adult Learners). Available online www.coe.int/T/DG4/Portfolio/documents/videoperform.pdf

Sigott, G. & Goetzinger, J. (2006). Report on analysis of rating data. Benchmarking Conference German: Spoken Performance, Vienna, 7-10 December 2006, organised by the ÖSD. (unpublished ms.).