Brian North
Zürich

Assessing Spoken Performance in relation to the Common European Framework of Reference

L'évaluation des compétences orales comprend deux aspects: a) la récolte d'éléments divers pour une performance et b) le jugement de la qualité de la performance. L'évaluation suppose une tâche qui permette de produire différents types de discours. Quelques variables-clé: production/interaction, thème reçu/thème choisi, présentation préparée/discussion spontanée, interaction avec l'enseignant/avec un autre élève.

Des exemples d'une approche particulière dans l'esprit de "bias for best" et tendant à l'autonomie sont illustrés dans un DVD pour l'anglais et un pour le français produits par le CERL. En effet, un des problèmes majeurs de l'évaluation de l'orale est lié à la difficulté d'établir une interprétation commune des critères. Il s'agit donc de vérifier que les examinateurs utilisent réellement ces critères et ne se basent pas sur une vision subjective des niveaux.

L'article présente des stratégies pour surmonter ce type de difficultés ainsi que des moyens pour comparer différents matériels d'évaluation régionaux.

Perhaps the major problem in oral assessment is assuring that there is a common interpretation of the assessment criteria concerned. With regard to assessment in relation to the Common European Framework of Reference (CEFR) (Council of Europe 2001), this means a common interpretation of the CEFR levels.

1. CEFR Standardisation Videos

To assist in achieving a common interpretation of the levels and criterion descriptors when assessing spoken performance in relation to the CEFR a DVD for French has been released this summer by the CIEP and Eurocentres (Lepage et North 2005a). This DVD is the product of the first international benchmarking conference in relation to the CEFR held in Sèvres in December 2005 for the Council of Europe by the CIEP and Eurocentres and for which the report (Lepage and North 2005b, 2005c) is available with documentation to the DVD on the Council of Europe website www.coe.int/lang. One third of the 38 participants at this seminar were from the CIEP and Eurocentres France, whilst the others were from the Alliance Française and other language schools in France, plus French specialists and CEFR specialists from around Europe. The DVD provides 23 calibrated spoken performances by young adults of 14 nationalities at all levels from below A1 to C2. A video for English of samples calibrated to the CEFR was also produced by Eurocentres and the Migros club schools in late 2003 (North and Hughes 2003) based on recordings of Swiss adult learners made during the research project that produced the CEFR levels and descriptors (North and Schneider 1998; North 2000; Schneider and North 2000). Information on the French and English videos can be obtained from bjnorth@eurocentres.com or Johanna.Panthier@coe.int. DVDs for German and Italian will appear early next year as a follow-up to seminars being organised by the Goethe Institute and the University of Perugia on the Sèvres model taking place in late 2005. Plans for a Spanish DVD also exist. Information on progress with the German, Italian and Spanish DVDs can be obtained from Waldemar.Martyniuk@coe.int.

2. Assessment in Relation to the CEFR

The assessment of spoken language proficiency consists of two aspects: (a) the elicitation of a sample of performance and (b) the judgement of the quality of that performance. In order for any comparisons to be made between the results obtained in the assessment and in order for any generalisations to be made between those assessment results and the results obtained by other candidates on other assessments, some form of standardisation as regards both the tasks performed and the interpretation of the criteria is necessary. Clearly there are limits in the extent to which such standardisation is achievable or desirable in a European context. The CEFR is *not* an exercise in European harmonisation; it is a tool intended to help language professionals to reflect on their current practice and to adopt a similar metalanguage in order to communicate with each other about the decisions they take. Those decisions should be defined by the context; the CEFR is not in any way intended to prescribe what the results of those decisions should be.

However, if one wants to talk about "assessment in relation to the CEFR," then it would be sensible to take account of certain principles to be found in the CEFR in the design of both the assessment task and the assessment criteria. These include issues like the following.

- **Transparency**: Informing learners about the learning objectives and assessment criteria to an extent appropriate to their age and experience. This might extend to self-assessment, but should at least include sensitisation to the nature of the task and to the different qualitative aspects of a good performance.
- **An action-oriented approach**: Assessing performance in activities that reflect what the learners will need do in the language. So, for example, no role plays: "persuade your parents to allow you to ....."[1].
- **Communicative language activities**: Ensuring the inclusion of phases of both spoken production (sustained monologue) and spoken interaction (spontaneous dialogue).
- **Communicative language competence**: Assessing pragmatic factors (e.g. fluency, precision, coherence) as well as linguistic factors. Valuing knowledge (range) as well as control (accuracy) when assessing linguistic competence.

3. Eliciting an Adequate Spoken Sample

An assessment of spoken proficiency requires a task that has phases that elicit different types of discourse. Some key variables are: production or interaction; given themes or chosen themes; prepared talk or spontaneous talk; colleague as interlocutor or teacher as interlocutor.

In the spoken assessment tasks used in the project that produced the CEFR levels (ibid.) and in the production of the calibrated samples produced for the Council of Europe for English and French, the following task was used. A prepared production phase by each candidate on a chosen theme (with 10 minutes to reflect beforehand on what to say) was followed by the other candidate asking questions at the end. After the second speaker's production phase, the pair drew cards with discussion topics. They were allowed to discard themes they didn't like and to go on to a new theme when they had nothing more to say, so again had choice of theme. The interaction phase was 100% spontaneous. The entire activity took 12-15 minutes.

Bambini che ascoltano una storia.

That is one particular style of elicitation that leaves the learners a great deal of autonomy to show their best. It was adopted because it is very simple to set up in classroom contexts, because it avoids the complications of the examiner being a part of the performance they are evaluating, and because it combined the spirit of "bias for best" and autonomy inherent in the Portfolio. During the recordings for the French DVD, the candidates were also recorded doing the DELF/DALF tasks for the new version of the exams to be released early next year. These activities are also very natural and almost invariably concern the candidate explaining the information from a "texte déclencheur" to a native speaker examiner, who then probes with follow-up questions. In one case a (Serbian, male) candidate performed noticeably better on a DALF task of this type than in the more relaxed atmosphere with a colleague as interlocutor. On the DALF sample he was (after statistical analysis) placed at the top of the C1 range and on the production/interaction sample he was placed at the bottom of the same C1 range. In another case a (Chinese, female) candidate performed far better on the informal task in which she managed the interaction in a very sophisticated manner, showed some C1 features and was calibrated at B2+, whereas she was only B2 on the formal DELF task in which she felt at a disadvantage to the interviewer. (The interviewer was Sylvie Lepage, to whom she then chatted happily over lunch and there was no doubt that her better performance on the production/interaction task was more representative). With other candidates, the differences in performances on the formal task (with interviewer) and informal production/interaction task (with colleague) were less marked. Nevertheless, as a result of this experience, for the filming of tasks for the German DVD by the Goethe Institute, a short phase with a native-speaker interlocutor has been inserted at the end of each production phase. Thus the examiner will ask questions at the end of this phase as well as or instead of the colleague.

There are, or course, many other approaches to elicitation. Cambridge ESOL, for example, focus on defining tasks for phases that elicit different types of discourse, that can be explained easily to candidates, that can be mass produced and yet guaranteed to be comparable, and that can be reliably carried out in a standardised fashion. The result is slightly reminiscent of a TV quiz show: very clear rules, an authenticity created by the very acceptance of the unnaturalness of the situation and a lot of talk from the "moderator." Cambridge samples for the main ESOL examination suite identified with levels A2-C2 are available on another standardisation video made for the Council of Europe (information from Johanna.Panthier@coe.int or direct from Cambridge ESOL).

4. Assessment Criteria

There are two fundamental choices to be made in the adoption of assessment criteria. Firstly, do the criteria relate to the specific task being performed, or are they "generic" in that they evaluate more abstract qualitative factors reflecting underlying communicative language competence? Fulcher (1996) argues very cogently that for any result to be generalisable, the latter approach must be taken. In the context of assessment in relation to the CEFR the obvious generic criteria to adopt are formulations taken from or adapted from those in the descriptor scales in CEFR Chapter 5. These define different qualitative aspects of language proficiency. They are summarised in CEFR Table 3 (English pages 28-29, French page 28, German pages 37-38). This table uses descriptors from Chapter 5 to define Range (Etendue), Accuracy (Correction), Fluency (Aisance), Interaction (Interaction) and Coherence (Cohérence) for each CEFR level. Pronunciation was excluded from CEFR Table 3 because it was designed for use in contexts in which assessors with different mother tongues might assess candidates of different mother tongues. In such contexts, assessment of pronunciation can be inconsistent, particularly since foreign-sounding pronunciation is usually thought of and defined in negative terms – more is bad (North 2000: 239-242). A descriptor scale for phonological control is in fact provided in the CEFR (English page 117, French page 92, German page 117) but, in common with the scale for orthographic control, it does not have the same high degree of validity as the other CEFR descriptor scales.

If all the candidates are at a particular range of level (e.g. around B1) then it would make sense to use an adapted version of CEFR Table 3 focussed on those and adjacent levels, perhaps including descriptors for the "plus levels" defined in the project that produced the descriptors (North and Schneider 1998, North 2000).

The second decision is: does one always use the same criteria (e.g. the 5 criteria of CEFR Table 3) or should one assess with the (say) three most relevant criteria for the task at hand? CEFR Table 3 was designed to be used with a single assessment task including both production and interaction phases. But if, for example, interaction and production tasks were going to take place at different times, and if the production task were to be a short talk, then it might make sense to rate the short talk with: Range, Accuracy, Precision and Coherence, whilst the interaction tasks might be rated with Range, Accuracy, Fluency, and Interaction.

5. Assessment Procedure

One of the biggest problems with oral assessment is raters not using the criteria, but assessing in relation to a personal view of the levels that they have developed independently. It therefore helps to have a phase of the assessment procedure in which the assessor consciously reads the criteria to check if the view of the performance that they are in the process of forming is in fact justified by the criteria. A procedure used for over a decade in Eurocentres (North 1993) and recommended in the Manual for relating examinations to the CEFR (Council of Europe 2003, Figueras et al 2005) can be summarised as three steps:

1. **Impression**: Write down the overall impression of the global level of the candidate that you have after about 3 or 4 minutes.
2. **Analysis**: Consciously read the descriptors for that level across the assessment grid. If you confirm that the candidate does meet the criterion description for a category at that level, look at the level above in that same category to see if they are even better than that. Write a result for each assessment category (Range, Accuracy, Fluency, Interaction, Coherence if using CEFR Table 3).
3. **Judgement**: Compare your analysis result to your original impression and make a considered judgement. Consult the CEFR scales for "Overall Spoken Interaction" and "Overall Spoken Production" if you find it difficult to make a final decision.

6. Alternatives

Clearly what has been described above is only one of many possible ways of eliciting and judging a spoken performance in relation to the CEFR levels. Many teachers will prefer to develop their own criteria and to use these to assess activities that are more related to their own curriculum.

6.1. Benchmarking Local Samples

In cases such of these in which no CEFR instrument is used in the operational assessment process, then in order to assure linking to the CEFR levels, one could follow the procedure outlined in the Manual for relating exams to the CEFR that was referred to earlier. This Manual proposes three phases of linking, of which the second phase "Standardisation" is relevant to the current discussion. In order to benchmark a local oral assessment approach to the CEFR, the Manual recommends videoing performance samples from the local assessment activities that have also been graded with the local criteria. Then these can be related to the CEFR in a one- or two-day "benchmarking seminar" that follows the following stages:

1. **Familiarisation** activities designed to ensure that the participants have an in-depth understanding of the CEF levels.
2. **Training** of judgements using the documented, calibrated performances provided by the Council of Europe Languages Policy Division in order to ensure an interpretation of the CEFR levels comparable to the interpretation elsewhere.
3. **Benchmarking** the local samples by immediately applying the consensus gained to the assessment of the local samples with the same criteria (CEFR Table 3).

6.2. Assessing and Benchmarking Discrete Tasks:

The technique discussed in this article is more suitable for levels at which learners are able to sustain an interaction or production. For the earlier years of lower secondary, assessment of the performance of tasks described by individual Portfolio descriptors may be more appropriate. One could imagine a simple assessment grid containing just 3 or 4 descriptors from the Portfolio checklist for the level concerned: the one(s) describing the task(s) being performed, plus those for different "Qualities" provided from Level A2 upwards. In relation to this approach the Dutch examination authority Cito are making available for benchmarking sessions a collection of mpeg video recordings for English of learners performing tasks from the Dutch secondary school Portfolio. For the second, Training phase of a benchmarking seminar such as that outlined above, these calibrated 2-3 minute extracts could be assessed onto CEFR levels in a benchmarking seminar using the holistic scale provided in the Manual as Table 5.8. Then the local Portfolio samples could be benchmarked with the same scale. More information about the Dutch benchmarking samples can be obtained from José Noijons at Jose.Noijons@citogroep.nl.

Footnote

[1] Unfortunately this is not an invented example.

References

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 (CEF). Preliminary Pilot Version of a Proposed Manual.* Strasbourg, Council of Europe, Languages Policy Division, DGIV/EDU/LANG (2003) 5. September 2003.

FIGUERAS, N. / NORTH, B. / TAKALA, S. / VERHELST, N. / VAN AVERMAET, P. (2005): *Relating Examinations to the Common European Framework: a Manual.* Language Testing, 22, 3, p. 1-19.

FULCHER, G. (1996): *Testing tasks: issues in task design and the group oral.* Language Testing 13, 1, p. 23–52.

LEPAGE, S. / NORTH, B. (2005a): *Exemples de productions orales illustrant, pour le français, les niveaux du Cadre européen commun de référence pour les langues.* Paris, DVD, CIEP et Eurocentres.

LEPAGE, S. / NORTH, B. (2005b): *Seminar to calibrate examples of spoken performances in line with the scales of the Common European Framework of Reference for Languages*, CIEP, Sèvres, 2 - 4 December 2004, Strasbourg, Council of Europe, Languages Policy Division, DGIV/EDU/LANG (2005) 1.

LEPAGE, S. / NORTH, B. (2005c): *Séminaire pour le calibrage des productions orales par rapport aux échelles du Cadre européen commun de référence pour les langues.* CIEP, Sèvres, 2 - 4 décembre 2004. Strasbourg, Conseil de l'Europe, Division des Politiques Linguistiques, DGIV/EDU/LANG (2005) 1.

NORTH, B. (1993): *L'évaluation collective dans les Eurocentres*, in: *Evaluations et Certifications en Langue Etrangère, numéro spécial*, Le Français dans le Monde - Recherches et Applications, août-septembre 1993, p. 69-81.

NORTH, B. (2000). *The development of a common framework scale of language proficiency.* New York, Peter Lang.

NORTH, B. / HUGHES, G. (2003): *CEF Performance Samples: for relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment. English. Swiss Adult Learners.* Council of Europe, Eurocentres, Migros Club Schools. Video cassette.

NORTH, B. / SCHNEIDER, G. (1998): *Scaling descriptors for language proficiency scales.* Language Testing 15, 2, p. 217–262.

SCHNEIDER, G. / NORTH, B. (2000): *Fremdsprachen können: was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit.* Nationales Forschungsprogramm 33: Wirksamkeit unserer Bildungssysteme. Chur/Zürich, Verlag Rüegger.

Brian North

is Head of Academic Development at Eurocentres (www.eurocentres.com), a Zürich-based foundation teaching languages worldwide in countries in which they are spoken. He is co-author of the CEFR and editor of the Council of Europe's CEFR Manual for examination bodies. His PhD concerned the development of the levels and descriptors for the CEFR.