Fourth Annual EALTA Conference, 15th-17th June 2007
Symposium: The CEFR in Europe and beyond: Challenges and Experiences.

Response by Brian North

Tim **McNamara** might be surprised to know just how much the issue of the normative influence of the CEFR levels is discussed at Council of Europe events in Strasbourg. All European education ministries are concerned that the CEFR be adapted appropriately in their context. One shouldn't forget that the title of the CEFR is: "Common European Framework of Reference for Languages: Learning, teaching, assessment." Assessment is in third place; the language testing profession is a service industry to support teaching and learning. The CEFR is primarily a curriculum project and all countries in Europe are more interested in the potential of the CEFR to stimulate curriculum reform and hence improved teaching and learning than they are interested in harmonising levels. The CEFR provides a metalanguage to facilitate communication and common points of reference to encourage networking and exchange of expertise between language professionals. EALTA is a good example: the progress made in a mere 3 years (2007 conference of 250 people, Best Practice Guidelines in 32 languages) is astonishing. It is no exaggeration to say that this could not have happened without the CEFR.

A "common framework" like this is a social construct, a constructed consensus. The CEFR descriptors are scaled *shared perceptions* of proficiency – we do not actually know that those perceptions are **"correct"** and that language proficiency really is as depicted, but they present us with common reference points to discuss things. In the literature linking through a common framework is called "Social Moderation" and moderation does not work without standardisation. In my view, the process of fixing precisely what the CEFR levels mean cannot be separated from networking, training with the illustrative samples, and collaborative empirical studies. This is what members of EALTA are discovering for themselves. Nobody is forcing implementation of the CEFR, it is as Tim says a tool that has had a long, slow development over 20-30 years in the European modern languages world. It is part of the process in which Europe has grown from 7 to 45 countries in 50 years. There is no European Education Ministry pushing this. Basically, the language teaching/testing community in Europe have decided that they want to relate more to what each other are doing. No one is telling or encouraging people outside Europe to adopt the CEFR either. You are welcome to get involved, but it is your decision.

To turn to the projects we have had presented, the TestDaF and Taiwanese projects each apply the steps recommended in the pilot version of the Council of Europe's Manual for "Relating Language Examinations to the CEFR" to the best of their ability.

The TestDaF presentation demonstrates use of the Standardisation stage from the Manual to corroborate a content-based hypothesis that the grades reported from TestDaf span B2-C1. The successful corroboration is not entirely unconnected to the fact that the researchers report on Writing, where they are able to compare ***directly*** the CEFR illustrative samples, the criteria contained in the CEFR Writing Assessment Scale, and TestDaF samples. An empirical validation stage is planned.

The Taiwanese project team focus on Reading in their presentation, following all three Stages suggested by the CEFR Manual. The researchers' main concern is whether linking indirect tests is valid, given the lack of precision in the descriptors and the lack of clear guidelines. Here one must remember that the CEFR descriptors are intended to describe learner real-life behaviour, not test tasks. Descriptors for listening and reading will only be improved on the basis of research into the ways the interactions between different kinds of complexity contribute to difficulty, and the ways in which interactions between inherent skills and adopted strategies contribute to successful performance. As regards clear guidelines, how clear would it be appropriate for guidelines to be? To me the word "guidelines" suggests prescribed, step by step procedures. Surely we should stay with general principles:

- Specification of content
- Standardisation training
- Benchmarking
- Independent corroboration that the standard-setting worked (External validation)

The TOEFL project describes a different approach to the issue of identifying cut-scores for CEFR levels on indirect tests. The project used an ingenious, test-centred, judgement-based standard-setting procedure which, to my mind, raises two interesting aspects: firstly the fact that the project relies entirely on guesstimates by experts and secondly the fact that it has been carried out in isolation.

The method described relies solely on the guesstimation of item difficulty by judges. Yet different standard-setting methods used in isolation are known to produce different results; judged difficulty and empirical difficulty have a weak relationship and aggregating or discussing judgements does not remove the problem. The procedures described thus give a hypothesis as to the relationship between TOEFL scores and CEFR levels. In my view, an external validation study is therefore necessary to validate the provisional cut-scores as outlined in Section 6.3 of the CEFR Manual. ETS research reports document many such external validation studies relating TOEIC to the ILR scale through self-assessments, teacher assessments and interview results. Staying within the ETS research tradition, why not do the same type of correlational study with real data on TOEFL and the CEFR? Meanwhile, a claim based just on judgements is a weak claim for a "high-stakes" international test.

As regards the question of isolation, the CEFR and the CEFR Manual do not represent a "club." The CEFR Manual does not claim to state "the right way" to link tests to the CEFR, it offers guidance. At this Conference, we have heard from John De Jong about an effective way of linking to the CEFR (through exploiting the CEFR scale cut-scores on the logit scale from my research) that has nothing whatsoever to do with the procedures we outline in the Manual. ETS have an independent research tradition that is a lot longer than that of the CEFR and it is perfectly legitimate of them to do things a different way. However, without an external study, how do we know that the interpretation of the CEFR levels by the TOEFL judges bears any relationship to the consensus interpretation of CEFR levels, or to the interpretation that might have been formed with another TOEFL standard-setting group, even using the same method?

The only way to resolve these issues is collaboration and the collection of corroborative evidence. It would be nice to include in the standard-setting chapter of the revised edition of the Manual the method used by TOEFL to set provisional cut-scores. It would be great to work further on descriptors both from qualitative analysis of calibrated samples – once we all agree on the calibration. And what about a project to calibrate more CEFR descriptors? To me it is surprising that no one has seriously taken up the inclusiveness and open-endedness of the itembanking approach to descriptor development proposed in the CEFR research project, so that the CEFR illustrative descriptors remain those produced by two people in Switzerland 10 years ago. Such a project could address the question of the appropriateness of CEFR descriptors in different contexts through studies of DIF (Differential Item Functioning) – even outside Europe.