



Testing
& Evaluation
& Assessment



Putting the CEFR to Good Use

Edited by Judith Mader and Zeynep Urkun

Selected articles by the presenters of
the IATEFL Testing, Evaluation and Assessment Special Interest Group (TEA SIG)
and EALTA Conference in Barcelona, Spain
29-30 October, 2010

ISBN: 978-1-901095-37-1
Publisher: IATEFL (TEA SIG)
www.iatefl.org



Published by IATEFL
 Darwin College, University of Kent
 Canterbury, Kent CT2 7NY
www.iatefl.org

©IATEFL, Judith Mader, Zeynep Urkun and all the contributors

Copyright for individual papers remains vested in the contributors to whom applications for permission to reproduce should be made directly or with the assistance of the TEA SIG.

First published 2010

British Library Cataloguing in Publication Data
 Education
 Mader, Judith and Urkun, Zeynep (Ed)
 Putting the CEFR to Good Use

ISBN: 978-1-901095-37-1

CONTENTS

Editorial Notes	2
Information about the conference and thanks	3
Using the CEFR to Benchmark Learning Outcomes: a Case Study Simon Buckland	4
Intercultural Competence and the CEFR – What’s the connection? Rudi Camerer & Judith Mader	11
CEFR and contrastive rhetoric – what’s the link? Cecilie Carlsen	27
Putting the CEFR to Good Use – A Collaborative Challenge Gudrun Erickson	36
A Virtual Approach to CEFR at University Levels Isabel Herrando Rodrigo	44
A Critical Look at the CEFR “Phonological Control” Grid David Horner	50
Linking Certification to the CEFR: Do we need standard-setting? Brian North	58
Designing fair writing testing tasks Vasso Oikonomidou	74
Using the CEFR in the foreign language classroom Anne Dragemark Oscarson & Mats Oscarson	83
What to teach and assess from A1 to C1 Susan Sheehan	92
Putting the CEFR to Good Use: Activities and Outcomes in Finland Sauli Takala	96
Improving classroom assessment by using CEFR Sanja Wagner	106

- For information and hand-outs of the conference, please refer to www.ealta.eu.org
- For more information on the conference program, contact Zeynep Urkun at zeynepu@sabanciuniv.edu.
- For further queries regarding the content of articles, please contact the writers through e-mail.

Putting the CEFR to Good Use

This collection of papers marks the 5th TEASIG volume of contributions by presenters at TEASIG Conferences. The conference on Putting the CEFR to Good Use was jointly organised and held with EALTA in Barcelona in October 2010.

Since its publication in 2001, the CEFR has been used in a large number of different contexts and by different groups of learners, teachers and assessors, both within and outside Europe. The papers here reflect this range of uses and provide information on and insights into a variety of best practices. This range means that almost everyone involved in the training and testing of language and communication skills will find something of interest. The subjects of the contributions cover the use of the CEFR in different national contexts as well as for different age groups and particular skills.

They have not been grouped in any particular categories here as the manageable number means that the reader will be able to identify and choose those easily which apply to a particular context. There are certainly many insights to be gained from the wide experience of the many eminent authors reflected in the contents of their papers. Contact details are given so that projects and references can be followed up.

We would like to thank all those who helped to make the conference in Barcelona such a success and who are too numerous to be mentioned here. Particular thanks go to those presenters who submitted papers for this volume and provide readers with in-depth information on the subjects of their presentations.

All those who attended the conference will appreciate the opportunity to revisit the topics of the presentations and workshops they attended as well as find out more about those which they could not attend. For those who were not able to be at the conference, we hope that you find this overview interesting and edifying and that it will encourage you to attend one of the future TEASIG conferences.

Editors:

Judith Mader



Zeynep Urkun



Information about the conference and thanks

IATEFL Testing, Evaluation and Assessment Special Interest Group (the TEA SIG) and EALTA (European Association of Language Testing and Assessment) jointly organized a two-day conference in Barcelona on October 29 & 30, 2010, titled "Putting the CEFR to Good Use".

TEA SIG and EALTA had long been seeking to join forces for an event that would cater for the needs of the testing community at large and the conference bore fruit to the articles contained in this publication. Our starting point was to find a theme which would have a common appeal to European testing audiences and the final consensus was: what better theme than the CEFR?

The impact of the Common European Framework of Reference for Languages (2001) and of the Manual for Relating Examinations to the CEFR (2009) in the field of language testing and assessment has resulted in a growing number of research programs, linking projects and training endeavors. However, different constituencies in different contexts of use with different resources create different scenarios which require tailor-made approaches in terms of improving current practice, and managing change, competence building in the area of using the CEFR not only in exam contexts but also in classroom assessment, increasing the quality of test development and test administration procedures, developing procedures that guarantee transparency and accountability, encouraging the development of both formal and informal national and international networks








The aim of this conference was to look at how professionals in the field have addressed these issues, and to exchange ideas on how different constituencies can cooperate in order to improve testing and assessment practice(s) in Europe.

The conference was very well-attended by some 120 delegates from 23 different countries and consisted of 4 plenary talks, 16 selected talks and 4 workshops. Participants were able to watch these presentations, to get to know the beautiful city of Barcelona and its surroundings, enjoy the lovely weather and to communicate with other delegates.

This publication of IATEFL TEA SIG – EALTA Conference consists of the selected articles written by 11 presenters of the conference.

Our thanks go to all IATEFL and EALTA officers for their support and help, in particular Zeynep Urkun, Neus Figueras and Gudrun Erickson. We would like to express our gratitude and thank all the article contributors for such interesting, well-prepared content, as well as APAC (Associació de Professors d'Anglès de Catalunya) for their invaluable support in taking care of the many local organizational details, in particular Miquel Breton, the treasurer. Special thanks to our plenary speakers, Gudrun Erickson, John H.A.L de Jong, Brian North and Sauli Takala for supporting the conference by appearing as plenary speakers despite their extremely busy schedules.

IATEFL TEA SIG COMMITTEE

						
Dave Allan Publicity & Contacts	Carel Burghout Webmaster	Doris Froetscher Discussion Moderator	Sue Hackett Coordinator	Judith Mader Newsletter Editor	Carol Spoettl Events Coordinator	Zeynep Urkun Events Coordinator

EALTA EXECUTIVE COMMITTEE

					
John de Jong President	Gudrun Erickson Secretary	Nigel Downey Treasurer	Evelyn Reichard 2010 Conference Organiser	Claudia Beccheroni 2011 Conference Organiser	Angela Haselgren Chair of the membership committee

Using the CEFR to Benchmark Learning Outcomes: a Case Study

Simon Buckland

Wall Street Institute International, London



Introducing Wall Street Institute

Wall Street Institute (WSI), part of the Pearson Group of companies, is a global provider of English language teaching to adults, with over 170,000 students in 440 learning centres in 30 countries. WSI uses its own proprietary Blended Learning system, which integrates computer- and print-based guided self-instruction with face-to-face small group classes led by teachers. It also comprises face-to-face support provided by L1-speaking personal tutors, and an online learner community.

The Wall Street Institute curriculum

The Wall Street curriculum, which was originally introduced in the early 1980s, and which has been updated five times since then, is based on the Council of Europe 'Threshold Level' (Council of Europe, undated) – a precursor of today's Common European Framework of Reference (Council of Europe, 2001). The Wall Street curriculum is divided into 17 levels, going from zero beginner to approximately C1 on the CEFR. A diagram later in this document shows how they map to CEFR levels.

The Wall Street Institute blended learning system: objectives and benefits

The delivery platform has evolved over the years that it has been in use, from book and tape through CD-ROM to today's internet/intranet version – but the methodology employed has remained essentially the same, based on the following principles:

- A mixture of guided self-access instruction and teacher-led classes
- The guided self-access instruction takes place on computer, either in a learning centre (intranet) or online (internet)

- The teacher lessons use small groups: there are 3 different types, ranging from a maximum of 4 to a maximum of 12 students
- The curriculum is modular and sequential: where students progress to the next module after completing the current one.
- Both formative and summative assessment are built into the system, with a mixture of computer-based tests, learner self-evaluation and mostly task-based teacher assessment.

The system is intended to assure the following advantages for learners:

- They have unlimited access to asynchronous components when, where, and as long as they wish, with no need to book lessons and no possibility of missing them.
- Students may book classes with teachers at any time that is convenient for them. To ensure this, lessons are scheduled on a recycling timetable and repeated several times a week or month, depending on demand¹.
- Learner monitoring is constant (and both summative and formative), with formative assessment by teachers of learners' progress and performance as well as test- and exam-type online assessment
- Instant progress feedback is available to students (in the form of a graphic profile) and full learner data for use by academic support and guidance staff. This is made possible by online storage of all study records (both synchronous and asynchronous); this student database is instantly accessible from any WSI learning centre in the world.

Challenges to standard-setting faced by the WSI system

The advantages of a blended learning system such as that used by WSI are fairly well established (Garrison and Kanuka, 2004; Rossett and Frazee, 2006 – among many). But its use in the context of Wall Street Institute also poses a number of challenges as regards standard-setting and quality assurance.

- The Wall Street network is highly distributed and localised, both globally and nationally (with as many as 60-70 centres in some countries)
- This makes it highly important to be able to compare the performance of one country with another, and one centre with another in the same country
- Accountability to customers is a fundamental commercial and ethical principle for any corporate solution provider
- Customers requiring measurable ROI (return on investment) also require evidence of the

¹ Very popular levels such as mid-A2 are scheduled much more frequently than, say, zero Beginner or C1 level classes.

efficacy of the learning system

- For commercial as well as homologation purposes, it's highly beneficial to conform to international norms (for instance, WSI is also recognised as compliant with the ISO9000 standard for service organisations).

Aims of the alignment study

For the above reasons, Wall Street Institute decided in 2006 to launch a formal study to verify and validate the alignment between our internal levels and the Common European Framework.

The aims that we set ourselves were as follows:

- to demonstrate recognized and accredited learning outcomes to students and those paying for their courses
- to provide verifiable benchmarking for students and their (potential) employers
- to comply with current and future requirements of governments and other regulatory bodies

The chosen route towards achieving those aims was to determine the match of the WSI curriculum and of students' learning outcomes with the CEFR.

The alignment study and its two phases

The entire project was carried out under the direction of Dr. Tony Lee, formerly the Head of Language Testing for the Hong Kong Government, and in association with Dr. Ardeshir Geranpayeh of Cambridge ESOL.

In order to 'triangulate' the objective as far as possible, the alignment study was divided into two phases: a quantitative and a qualitative phase.

Phase One (quantitative)

The chosen instrument for Phase One was the Cambridge BULATS test (the computer version), which has already been correlated to the CEFR by Cambridge ESOL themselves (BULATS Research, online). Given that we already know which Wall Street level our students are at, the goal was to map these onto their BULATS scores, and thence onto CEFR levels.

A total of 5,688 students were tested in six different countries: Argentina, Germany, Hong Kong,

Italy, Portugal, Saudi Arabia. These were chosen in order to reflect a representative range of L1s and cultural backgrounds. The results were then calibrated using Rasch models (Bond & Fox, 2007).

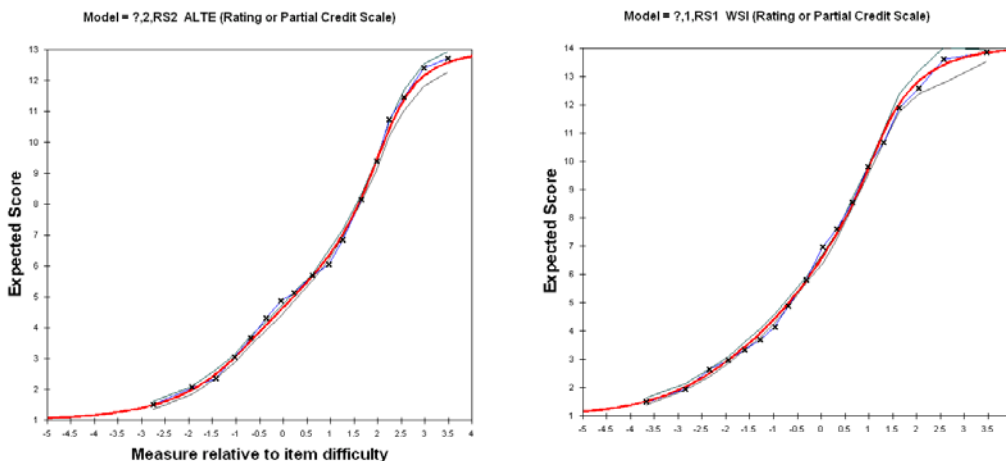
Phase Two (qualitative)

The aim of Phase Two was to align WSI curriculum content with the CEFR, following a similar methodology to that employed in the development of the Framework of Reference itself. A random selection was made of about 200 ALTE/CEFR Can-Do statements, and placed into a random sequence selection, with the CEFR level removed. An expert panel of 90 WSI teachers with 18 months or more of work experience at WSI was given this selection of Can-Do statements and asked to match them to WSI levels. Again, the results were Rasch calibrated.

Results and conclusions of the alignment study

The results of the alignment study conformed to WSI's expectations, with a high degree of correlation in both phases.

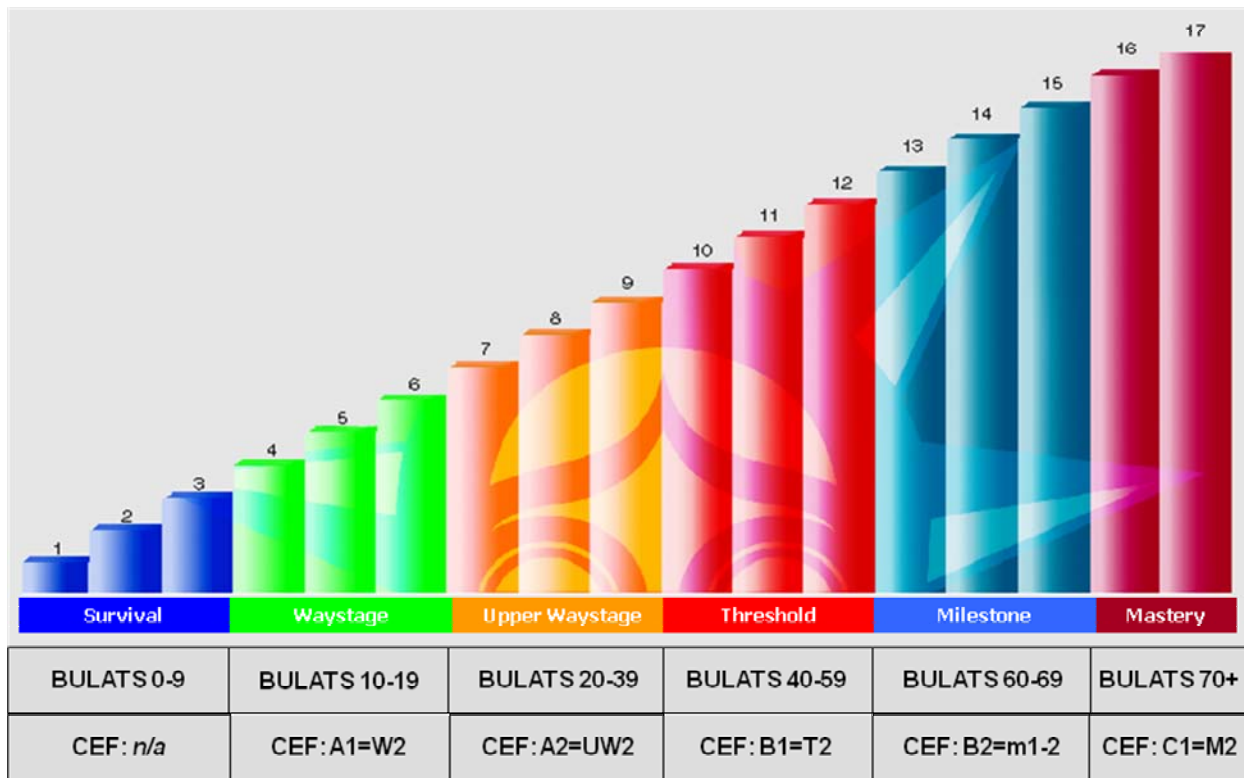
In Phase 1 there was a good to excellent match between WSI levels and BULATS scores, showing that the former are effective discriminators (see diagram).



Using the alignment which Cambridge ESOL have demonstrated, WSI are thus able to link WSI internal levels to Common European Framework levels.

In Phase 2 we found a more than adequate match between the WSI raters' assessments and the CEFR levels – almost 80% of the Can-Do statements had a good or high level of correlation.

Our overall conclusion was that the link between CEFR levels and WSI's internal levels had been satisfactorily established. The study was conducted in close consultation with Cambridge ESOL, who have formally and publicly endorsed the methodology employed. On this basis, WSI have published the results of the alignment study for use in our commercial and educational communications with students and prospective students (see below):



The study was presented at the ALTE and IATEFL conferences in 2008, and published in the proceedings of those conferences (Taylor & Weir, 2009 and Beaven, 2009).

As Cambridge ESOL acknowledged in their public statement, Wall Street Institute is the first global ELT provider to carry out such a study. Our hypothesis is that the blended instruction and assessment model enables WSI to achieve a higher level of consistency across countries and learning centres than is possible using a purely instructor-led model.

Using the alignment in practice

The goal of the study, as explained, was to provide a benchmark and quality guarantee to actual and prospective customers, along the lines of the ISO9000 certification which WSI has also

obtained. However, there have been a number of challenges in achieving this goal.

1. The concept of alignment is somewhat “academic”, and course consultants find it hard to explain to non-experts
2. The CEFR is not well known outside a few European countries (it is by no means universally familiar even across Europe).
3. The world at large still doesn’t really understand the difference between assessing ability using can-do statements and assessing knowledge. To put it another way, the world in which WSI operates remains highly exam-focused – and not always on the same exams at that (there are “Cambridge countries”, favouring FCE, IELTS etc. and “ETS countries”, favouring TOEFL and TOEIC).
4. Last but not least, unsubstantiated and unreasonable claims by competitors (see below for an example) muddy the waters, and make it hard for customers to assess the merits of our claims and of the work that underlies them.



Improbable claims made by Berlitz for the alignment of their courses with the CEFR. Each of the 10 columns represents 80 hours of instruction, and the diagram suggests 320 hours from the start of B1 to the end of B2, but only 80 hours for the whole of C!

Conclusion and discussion

Although WSI was satisfied with the execution and results of the alignment study – and we intend to repeat the exercise in 2011 – we felt that it raises some interesting questions. I leave them for readers to reflect on and possibly discuss – and I will be happy to do so by email with anyone who is interested:

- What does “alignment to the CEFR” actually mean?
- Are you “aligned” if you teach courses leading to the Cambridge Main Suite exams?
- What if your students take TOEFL and TOEIC – are you equally “aligned”?

- How do you “prove” alignment to the CEFR?
- Can there ever be an ISO-type auditing process?
- Can there ever be sanctions against false claims?
- What is the market looking for, can it be provided and if so, how?

References

Beaven, B. (ed.) (2009) *IATEFL 2008 Exeter Conference Selections*. IATEFL

Bond, T.G. & Fox, C.M. (2007) *Applying the Rasch Model – Fundamental Measurement for the Human Sciences*, 2nd ed. Lawrence Erlbaum: New Jersey, London

Cambridge ESOL, University of Salamanca, Goethe-Institut, Alliance Française and ALTE: BULATS Research. <http://www.bulats.org/Bulats/research.html>

Council of Europe (2003) *Reference Level Descriptions (RLD) for national and regional languages*. http://www.coe.int/t/dg4/linguistic/DNR_EN.asp

Council of Europe (2001) *The Common European Framework of Reference for Languages – teaching, learning, assessment*. Strasbourg: Language Policy Division

Council of Europe (undated) *Threshold Level and CEFR*. http://www.coe.int/t/dg4/linguistic/DNR_EN.asp#P30_4262)

Garrison, D.R. & Kanuka, H. (2004) *Blended learning: Uncovering its transformative potential in higher education*. In: *The Internet and Higher Education*, Volume 7, Issue 2, 2nd Quarter 2004, 95-105

Rossett, A. and Frazee, R.V. (2006) *Blended Learning Opportunities*. American Management Association. <http://www.amanet.org/training/whitepapers/Blended-Learning-Opportunities-45.aspx>

Taylor, L. and Weir, C. (2009) *Language Testing Matters: Investigating the wider social and educational impact of assessment*. In: *Proceedings of the ALTE Cambridge Conference*, April 2008. Cambridge: Cambridge University Press

sbuckland@wallstreetinstitute.com

Simon Buckland is the Director of Curriculum Development for Wall Street Institute International. He has been developing applications for computer-assisted training and language learning for many years, and has been the chief designer and developer of the Wall Street Institute blended learning system since the 1980s.

Intercultural Competence and the CEFR – What’s the connection?

Rudi Camerer, etc – European Language Competence, Frankfurt &

Judith Mader, European Language Competence & Frankfurt School of Finance and Management, Frankfurt



The CEFR – an action-oriented approach

The Common European Framework of Reference for Languages (CEFR) is “non-dogmatic: not irrevocably and exclusively attached to any one of a number of competing linguistic or educational theories or practices.”² This non-dogmatic approach is underlined by the constant reminders that “Users of the Framework may wish to consider...” and is one of the many strengths of the CEFR, strengths which have undoubtedly led to its widespread acceptance in large parts of Europe and beyond. The non-dogmatic approach of the CEFR does not however mean that it is without a standpoint, as is made clear by the authors in their insistence on language as a means of communication. As with a constructivist view of culture, it is not what language **is** that is important in and for the CEFR, but what language **does**. This understanding of language is further clarified, reinforced and exemplified by the 54 descriptive scales, of which only 4 have linguistic accuracy as their focus – grammar, vocabulary, pronunciation and orthography. A yet unanswered question is that of the relation of these scales to each other. Does a hierarchy of language skills / can-do’s exist? Are some of the scales more important for successful communication than others? There is no descriptive scale for intercultural competence as such. Is there a (good) reason for this?

The declared intention of the Council of Europe (CoE) in producing and promoting the CEFR was to encourage and further plurilingualism in Europe and with it mobility and mutual

² CEFR p. 8

understanding.³ There can be little question that intercultural communication, and so logically intercultural competence, must play a role in the achievement of these aims. The language policy aims of the CoE and the aims of the CEFR described in the first chapter confirm this assumption.

The bringing together of mental and intellectual abilities on the one hand and practical communicative competence on the other is reflected in the discussion of intercultural competence in all the relevant disciplines⁴. However the claim of the CEFR that intercultural competence can be viewed in terms of active language use raises several questions, the answers to which have far-reaching consequences for teaching and assessment. These questions can be posed as follows:

- I. If language competence and intercultural competence belong inextricably together, they cannot be the same thing. How do they overlap, what do they have in common and in what ways are they different?
- II. How far does personality play a part in successful intercultural communication? Can features of personality be taught or – dare we say it – tested?
- III. How can intercultural communicative competence be described?
- IV. How much knowledge is necessary to be interculturally competent?
- V. Is there any sort of progression in the acquisition of intercultural competence, similar to that in the process of acquisition of language?
- VI. Which English is used in intercultural encounters?
- VII. Can intercultural competence be tested?

This paper attempts to go some way towards answering these questions one by one, referring to a project in which training and test material has been developed and used in practice in the context of training programs for chambers of commerce and industry in Germany and Austria.

I. Language competence and Intercultural competence: If language competence and intercultural competence belong inextricably together, they cannot be the same thing. How do they overlap, what do they have in common and in what ways are they different?

This question is closely linked to the question of a hierarchy of the descriptive scales of the CEFR. We have no doubt that in most intercultural encounters, it is aspects of politeness and cooperation, reflected in actions such as turn-taking and compensating that are more important

³ CEFR pp.1-8

⁴ cf. the overview provided by Helen Spencer-Oatey and Peter Franklin (2009). Intercultural interaction: a multidisciplinary approach to intercultural communication (Palgrave Macmillan)

than, for instance, sustained monologue and creative writing. Depending on the context of use, some scales are clearly more important than others. For our purposes the scale for *Sociolinguistic Appropriateness* proved particularly useful and in particular the remarks of the authors on the difficulty of producing such a scale:

*The scaling of items for aspects of sociolinguistic competence proved problematic (see Appendix B). Items successfully scaled are shown in the illustrative scale below. As can be seen, the bottom part of the scale concerns only markers of social relations and politeness conventions. From Level B2, users are then found able to express themselves adequately in language which is sociolinguistically appropriate to the situations and persons involved, and begin to acquire an ability to cope with variation of speech, plus a greater degree of control over register and idiom.*⁵

Finding and collating all the relevant descriptors for a description of intercultural competence in the CEFR demonstrated to us quite clearly the view of the authors of the CEFR that intercultural competence requires a minimum level of linguistic competence in the language concerned. However it is also clear – in contrast to what is claimed by the authors of the CEFR – that features of intercultural competent communicators can be observed alongside other features of users at level B1, for instance in the scales *Repairing*, *Oral Interaction* and *Monitoring and Repair*. This observation proved to have practical consequences for the development of training courses and material. Learners at level B1 should be able to use simple but effective communicative strategies to make intercultural communication successful, providing they possess other competences. It is these other competences which make up the training program referred to here. It seems therefore possible to link intercultural competence and language competence at level B1. The nature of this link and its meaning in practical communication is discussed in the following.

II. Intercultural competence and personality: How far does personality play a part in successful intercultural communication? Can features of personality be taught or – dare we say it – tested?

The authors of the CEFR are – understandably – reluctant to define and discuss the role of personality in the context of education in intercultural competence.

Attitudes and personality factors greatly affect not only the language users'/learners' roles in communicative acts but also their ability to learn. The development of an 'intercultural personality' involving both attitudes and awareness is seen by many as an important educational goal in its own right. Important ethical and pedagogic issues are raised, such as:

⁵ CEFR p. 121

- *the extent to which personality development can be an explicit educational objective;*
- *how cultural relativism is to be reconciled with ethical and moral integrity;*
- *which personality factors a) facilitate b) impede foreign or second language learning and acquisition;*
- *how learners can be helped to exploit strengths and overcome weaknesses;*
- *how the diversity of personalities can be reconciled with the constraints imposed on and by educational systems.*⁶

There is no doubt that intercultural competence is a blend of competences, reduced to basically three – knowledge, willingness, and ability. The desire or wish to communicate intercultural will inevitably imply characteristics such as tolerance, openness or empathy. Whether these characteristics are specific to intercultural competence or whether they are features of communicative competence in general, i.e. also intraculturally relevant, would be the first question to be raised if it is specifically intercultural competence which is being discussed. The second and more important question is how far these characteristics can be taught or trained in adults. However much it may be the aim of school education to raise tolerant open-minded citizens, it is not always easy or even possible to change lifelong attitudes in adults. We have reservations as to whether this is possible at all and, if it is, whether it is something which language teachers should be trained to do. We are not talking here about aspects of language teaching such as learning to learn or counselling learners on learning, which may be part of language teaching for adults (for both learners and teachers) but we would be wary of expecting too much from or putting too great a burden on language teachers. This is also referred to by the CEFR authors and those who have commented on it who warn against using its approach for purposes of social engineering.⁷

The CEFR makes use of the five *savoirs*, suggested by Michael Byram (1997) in *Teaching and Assessing Intercultural Competence* and which claim to cover the entire spectrum of abilities, attitudes and knowledge required for successful intercultural communication.

⁶ CEFR p. 106

⁷ Heyworth, Frank (2004). Why the CEF is important. In: Morrow, Keith (2004) (Ed.). *Insights from the Common European Framework*. OUP. P. 14

	Skills interpret and relate (<i>savoir comprendre</i>)	
Knowledge of self and other; of interaction: individual and societal (<i>savoirs</i>)	Education political education critical cultural awareness (<i>savoir s'engager</i>)	Attitudes relativising self valuing other (<i>savoir être</i>)
	Skills discover and /or interact (<i>savoir apprendre/faire</i>)	

It is certainly extremely helpful for the development of curricula and teaching material to be able to base these on an accepted description of the various factors affecting intercultural competence. We assume however that it is performance which is the most important aspect and what really counts in intercultural communication. Questions of personality, what interlocutors/participants in the communication “really” think and feel, may well not appear on the surface, often remain concealed and may not be decisive for the success of the communication if those involved communicate successfully and achieve their goals.

Another reason for our reservations as far as features of personality are concerned, is that these cannot be evaluated or tested in any way which approaches the way good language tests are constructed. Even though the opposite is often claimed⁸, there are no valid, reliable and objective tests of personality available and these are not to be expected in the near future. We say more about this below.

III. Defining intercultural competence: How can intercultural communicative competence be described?

The CEFR places the focus firmly on communicative competence and describes this at six levels. Can intercultural competence at different levels be described and which descriptors would be necessary? There have been attempts to do precisely this and the success or failure of these attempts has played an important part in the development of training material for courses claiming to teach intercultural competence. In order to answer the question above, it is first

⁸ cf. the 53 tests / assessment tools / indicators etc. of intercultural competence, which are currently listed on the SIETAR-Europa website <http://www.sietar-europa.org/SIETARproject/Assessments&instruments.html#Topic26>

necessary to define the issue more closely. Intercultural competence is more than just language competence, there seems to be a general consensus on this. The question of which competences a linguistically competent person needs in order to be regarded as interculturally competent is still a matter of lively debate. Depending on the academic discipline concerned, these additional competences range from behavioural flexibility through emotional resilience to adaptation or integration. The difficulty is agreeing on a definition of any of these terms and concepts. Based on parts of the CEFR as well as the work of Meierkord (1996), Beneke (2000), Byram (1997, 2001), Müller-Jacquier (1999, 2000), Wolf and Polzenhagen (2006) and others, we would suggest the following criteria for intercultural competence. These describe the characteristics of an interculturally competent person in such a way to allow them to be used as test criteria as well as for the development of a test format, valid items and a practicable marking system. The following eight criteria are designed to relate to the active use of language in intercultural encounters and to be used for standardised, objective testing procedures.

These are:

1. **Knowledge about institutions, processes of socialisation and other specifics in one's own and in one or more target countries.** This includes country specific knowledge of one's own as well as the other culture(s) one may have to deal with. As well as being aware of and able to use appropriate discourse conventions (which we comment on below) it is also important for the success of intercultural communication that the interlocutors appear interested in and informed about the other's culture, rather than uninterested and ill-informed. A certain amount (the particular amount is to be defined in each particular case) of knowledge about specific countries / cultures is necessary. This may range from knowing what the local currency and the capital city are to being informed about social, economic, political or religious features of the country or culture. Typical patterns of behaviour (Dos & Don'ts) as well as information on local literature, music and art ("high culture") may also form part of this range of knowledge. It is by no means necessary to possess comprehensive information on all these aspects of culture. This criterion focuses on the awareness of the necessity to acquire a basic amount of this type of information.
2. **Knowledge of the causes and processes of misunderstanding between members of different cultures.** This implies awareness of and familiarity with the particularities of one's own as well as the other culture. Examples of potential critical cultural distinctions are the notions of time, hierarchy, space etc. Examples of potential critical discourse functions

are refusing / rejecting, contradicting, instructing, criticising, disagreeing, making and receiving compliments, complaining and dealing with complaints, among others. One's own personal and culturally-influenced discourse strategies should be the focus of critical appraisal as well as those of the other culture. A range of interculturally appropriate ways of dealing with these language functions should be available to language users.

3. **Ability to engage with differences in a relationship of equality (including the ability to question the values and presuppositions in cultural practices and products in one's own environment).** The most important feature of this criterion is the ability to express oneself non-judgementally or neutrally on culturally significant phenomena as well as on stereotypes. It also includes the ability, in intercultural encounters, to comment on generalisations with the required amount of objectivity without endangering the relationship to the interlocutor.

4. **Ability to engage with politeness conventions and communication and interaction conventions (verbal and non-verbal).** Politeness is the key term in and the key feature of intercultural communication. Politeness means more than simply following rules of etiquette (however important these may be) , it is more concerned with the building of positive relationships, particularly in first and second encounters. It is often in these encounters that the ground is laid for the nature of the relationship and its medium- or long-term success or failure. The ability to interact with the necessary degree of politeness in intercultural encounters is not easy to acquire, as politeness (discourse) conventions differ so greatly from culture to culture and language to language. What may be regarded as perfectly acceptable and appropriate in one culture may be totally unacceptable in another. Learners should therefore have some knowledge of the existing conventions (which apply to the particular encounter) as well as being aware of possible signals and reactions which may signify irritation, confusion or even anger. As well as perceiving these signals, it is also necessary to be able to deal with them in an appropriate way. Being able to use language politely to maintain and possibly repair relationships is therefore a key factor in intercultural communication.

5. **Ability to use essential conventions of oral communication and to recognise changes in register.** This follows closely the criterion described above. Use of inappropriate register in communication is one of the most frequent causes of intercultural

misunderstandings and breakdowns in communication. Being able to use polite discourse conventions correctly and appropriately to deal with all situations, in particular critical or potentially critical interaction must form one of the most important goals of any intercultural training course. This requires familiarity with conventions of communication which may be appropriate or inappropriate for use with interlocutors from different cultures. Examples of these are forms of address, directness / indirectness, face-saving strategies etc.

6. **Ability to use essential conventions of written communication and to recognise changes in register.** What is described above applies equally to written communication, with the important difference that spontaneous repair strategies cannot be used. This makes proof-reading for possible contraventions of polite discourse conventions extremely important for written intercultural communication.

7. **Ability to elicit the concepts and values of documents or events** (i.e. meta-communication). It may be necessary, especially when the encounter is threatening to become critical, but also before this happens, to discuss the particular discourse and other conventions which prevail in order to ascertain what these are and to reach agreement on which conventions are appropriate and acceptable in the particular encounter. This must be done without any appearance of superiority or arrogance on either side and should not lead to embarrassment for either party. Communicating about the communication itself can be extremely helpful even if both parties are prepared for the encounter and willing to use (temporarily) the conventions applicable in the other culture. Metacommunicative discourse strategies have not yet been the focus of language training, although the mastery of these may be crucial for the building of a positive relationship. The importance of this criterion is not diminished by the fact that in many so-called "high-context" cultures attempts at meta-communication may be rejected implicitly. Knowledge of this possibility and the ability to use metacommunicative strategies appropriately in the relevant situations are what is meant by this criterion.

8. **Ability to mediate between conflicting interpretations of phenomena.** Some of what is said in the CEFR about mediation seems inconsistent. This poses a dilemma, as in some places in the CEFR mediation is taken to mean translation / interpretation, in others the central meaning is that of mediation in intercultural contexts, which broadens and changes its significance and may lead to a different interpretation. The summary in the penultimate

chapter on curriculum development both in and outside school describes the spectrum of skills referred to in the context of mediation:

*“... it would be helpful if the ability to cope with several languages or cultures could also be taken into account and recognised. Translating (or summarising) a second foreign language into a first foreign language, participating in an oral discussion involving several languages, **interpreting a cultural phenomenon in relation to another culture, are examples of mediation** (as defined in this document) which have their place to play in assessing and rewarding the ability to manage a plurilingual and pluricultural repertoire.”⁹*

It is this definition we have used as the basis for the development of the curriculum and test in intercultural communicative competence.

These eight criteria are indicators of intercultural communicative competence. The ability to use each of them requires a minimum level of linguistic competence. Using the relevant CEFR descriptors as well as our own practical training experience and test piloting, we have set this minimum level at B1. The overriding aim of our extensively piloted training programme is to prepare learners to use their linguistic abilities in intercultural encounters in such a way that their communicative behaviour corresponds to the criteria described above as far as possible. The criterion-based test shows how far this goal has been achieved.

IV. Knowledge or ability: How much knowledge is necessary to be interculturally competent?

We have already referred to the importance of country- or culture-specific knowledge in intercultural communication. Learners should appreciate the relevance of this type of knowledge and not enter fully unprepared into intercultural encounters. However there is also a different kind of knowledge which may play a role in intercultural encounters. This is knowledge of the things which are apparently taken for granted in a particular culture and how these may differ from culture to culture, i.e. knowing that my view of reality may be different from that of my fellow interlocutor's, possibly even so different that there can be little or no mutual agreement.

Acquiring this knowledge does not imply a study of the literature or attendance at lectures, as is the case in many intercultural training courses.¹⁰ We believe that the uncritical acceptance of

⁹ CEFR p. 175 our emphasis

¹⁰ Jürgen Bolten, Interkulturelles Coaching, Mediation, Training und Consulting als Aufgaben des Personalmanagements

many so-called culture frameworks as described in detail by Hofstede, Trompenaars, Hampden-Turner and others may prove problematic. Learners should be prepared to encounter individuals rather than “a culture”. These individuals may – consciously or unconsciously – differ quite significantly in how they behave from the behaviour described as typical for their culture. In view of this, it is surprising how many training programmes present the findings of Hofstede as the content of “Intercultural Theory”, even though many other, more recent, data-based findings exist and are easily accessible. It is equally astonishing how criticisms of this approach are ignored. The differences between the findings of intercultural researchers and theorists are rarely mentioned nor are the variety of methodological approaches and the many questions which arise from this variety or the proposed number of cultural dimensions proposed.

Edward T. Hall 1950-60's	Geert Hofstede 1970's empirical research, IBM	Trompenaars / Hampden-Turner 1980's empirical research, Shell	European Values Study 1981 – 2004 ... empirical research	World Values Survey 1994 – 2008 ... empirical research	Globe Study 1991 / 2006 empirical research	Schwarz Value Survey 1999 empirical research
High context vs. low context communication		Universalism vs. particularism	Life	Perception of Life	Performance orientation	Conservatism
	Individualism vs. collectivism	Individualism vs. communitarianism	Society	Environment	Institutional collectivism	Intellectual autonomy
	Power distance	Neutral vs. emotional	Work	Work	Power distance	Affective autonomy
	Masculinity vs. femininity	Specific vs. diffuse	Family	Family	Gender egalitarianism	Hierarchy
Monochronic vs. polychronic	Short-term vs. long-term orientation	Sequentially vs. synchronically organised activities	Politics	Politics and Society	Future orientation	Egalitarianism
	Uncertainty avoidance	Controlling nature vs. letting it take its course	Religion	Religion and Morale	Uncertainty avoidance	Mastery
Space (proxemics)		Achievement vs. ascription		National Identity	Humane orientation	Harmony
					Assertiveness	
					In-group collectivism	

This leads to questions such as: Are there any number of culture dimensions? Are some more important than others? How are diametrically opposing results to be explained? How do any of the dimensions affect actual communicative behaviour in intercultural encounters? It is beyond the scope of this piece to attempt to answer these questions.¹¹

V. Levels of intercultural competence: Is there any sort of progression in the acquisition of intercultural competence, similar to that in the process of acquisition of language?

internationaler Unternehmen. In: J.Bolten / C. Ehrhardt(Hg.), Interkulturelle Kommunikation. Texte und Übungen zum interkulturellen Handeln (2003). Yvonne Knoll, Currently Offered Intercultural Training in Germany and Great Britain. An Empirical Study, in: Interculture Journal 2006/1

¹¹ cf. Judith Mader, Rudi Camerer: International English and the Teaching of Intercultural Communicative Competence. In: Interculture Journal 12/2010. pp. 97-116

The piloting of our curriculum and training programme indicate strongly that a progression exists. This progression in successful communication is not by any means to be equated with linguistic knowledge and ability. Successful intercultural competence does not principally depend on a high level of linguistic competence. The specific partial competences which make a linguistically competent speaker into an interculturally competent speaker are only partly dependent on linguistic knowledge and ability.

Although B1 is set at a minimum level, it can be assumed that users at B2, C1 and C2 have a greater range of discourse strategies (oral and written) at their disposal and may well be more confident in identifying differences in register or in the use of metacommunicative strategies. However the piloting showed that learners with relatively small linguistic means at their disposal were able, through repeated use of a small number of strategies, to communicate successfully in intercultural encounters. As well as this, we were able to establish that users at higher levels of linguistic competence did not demonstrate a significantly higher level of intercultural communicative competence. This may be connected to the fact that language users at a high linguistic level are often subject to stricter “rules” and less likely to be forgiven for intercultural “faux pas” than users who clearly do not possess the linguistic means to express themselves appropriately, for example less directly.

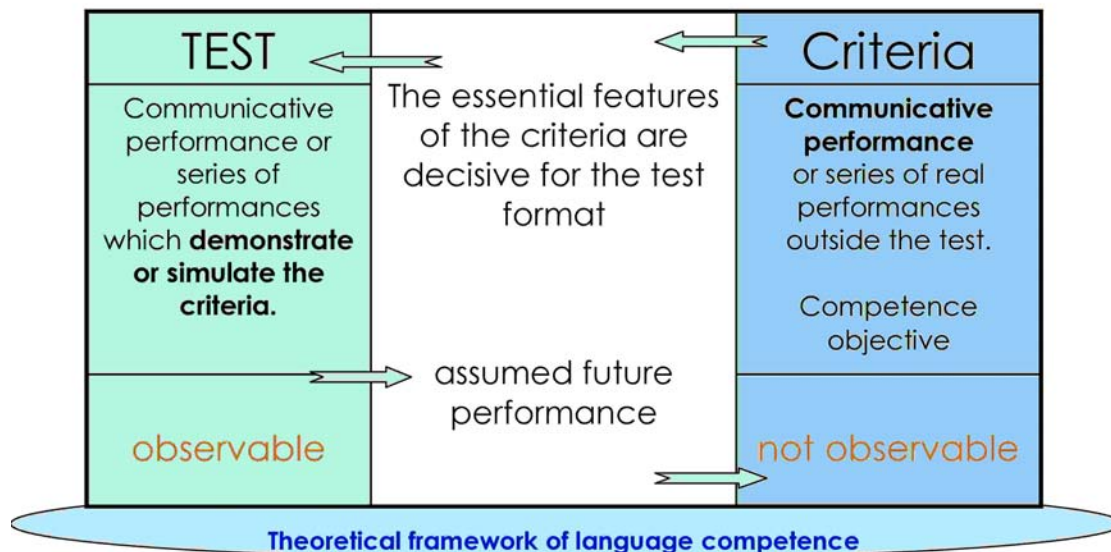
This leads to the question of whether intercultural competence can be trained and if so how? Being aware of the existence of different views of reality is undoubtedly significant for successful intercultural encounters. In contrast to many intercultural training programmes, we do not provide any input on culture dimensions and intercultural theory, but use learners’ own experiences to motivate them to reflect upon their own cultural standards and views of reality, as well as what they take for granted. Learners are also prepared to be confronted with entirely different views and values from their own and are provided with the appropriate discourse strategies to deal with these in specific situations. They are tested neither on their knowledge of cultural frameworks nor on their knowledge of specific intercultural theorists. Emphasis is placed on increasing learners’ familiarity with different cultural standards as well how these are expressed and on increasing their awareness of their own attitude to differences encountered and their own ways of behaviour, particularly language behaviour.

VI. Intercultural competence and English: Which English is used in intercultural encounters?

When set against the central aim of plurilingualism in European language policy, the fact that the CEFR first appeared in English can perhaps be seen as coincidental, particularly as it is now available in 32 languages, including some as far from “European”, as Chinese, Korean, Japanese or Arabic. Clearly some features of the CEFR are perceived as of common interest and importance by users of all these languages, a perception which has led to the acknowledgement of the CEFR in translation as well as, in many cases, to its adoption and official or semi-official recognition for educational purposes by institutions and bodies in Europe and beyond. The question of which specific variety of English to use cannot be answered here as it will clearly depend on the cultures to be encountered. Generally it can be said that the main varieties of English are British and US-American and standard variations on these as well as what is known as *Mid-Atlantic* and what is coming to be known as *International English / English as a Lingua franca*. We enter into the use of this in intercultural encounters as well as its implications for the test elsewhere.¹²

VII. Testing intercultural competence: Can intercultural competence be tested?

Tests of proficiency in a language should make possible a prognosis of future communicative behaviour, in this case in intercultural encounters. This can be represented as follows:



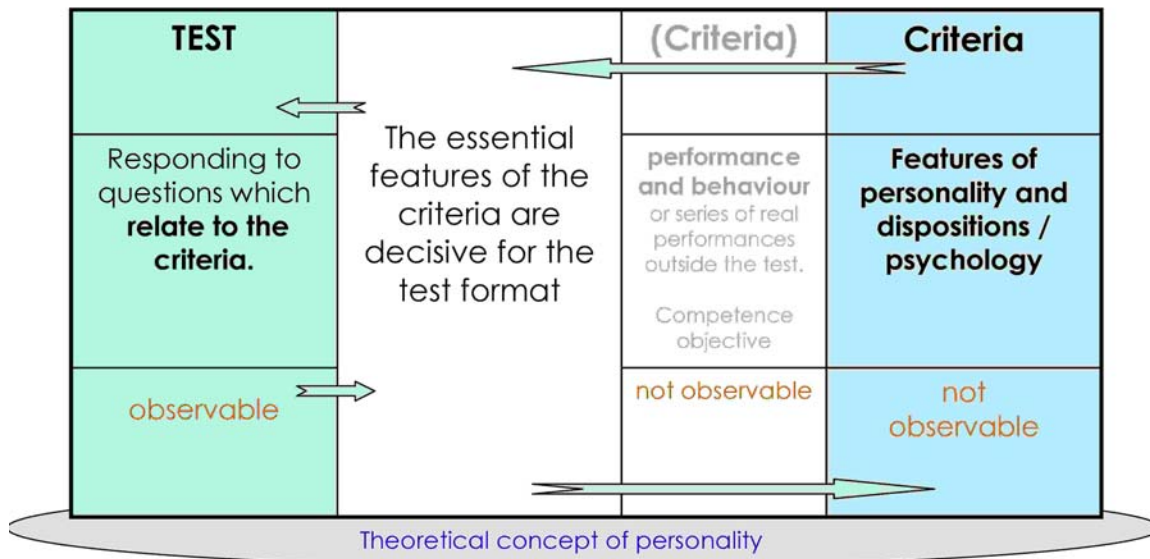
Read anticlockwise, the diagram shows the relationship between test and criteria.

¹² Camerer Mader Interculture journal

Essential features of any test construct are therefore:

- A description of the intended communicative competence (e.g. competence levels of the CEFR)
- A representative sample of partial competences
- A precise description of the marking criteria
- Standardised test procedures simulating these as well as standardised scoring procedures
- A test construct which bases its criteria on a comprehensive description of verbal, non-verbal and paraverbal communication, such as that to be found in the CEFR, can make a claim to validity which should meet with wide acceptance.

Tests of intercultural competence available on the market (such as the *Intercultural Development Inventory* developed by Milton J. Bennett and Mitchell R. Hammer and available worldwide or the *Test of Intercultural Sensitivity* provided by ICUnet AG based in Germany) provide a complete contrast to an approach of this type. These tests work with self-response questionnaires and the interpretation of these. The construct such as it is, can be portrayed as follows:



Although psychometric tests of this type are widely used in corporate personnel selection, the professional consensus on which they are based is minimal. Characteristics such as intelligence, aggression, attraction etc. are generally considered to be so abstract as to be largely irrelevant in connection with performance and the forecast of this. To underline this point, here is a quote

from a “sobering reminder about the low validities and other problems in using self-report personality tests”:

There is considerable evidence to suggest that when predictive validation studies are conducted with actual job applicants where independent criterion measures are collected, observed (uncorrected) validity is very low and often close to zero. This is a consistent and uncontroversial conclusion.¹³

A valid measurement of personality features and any empirical evidence of how they relate to practical performance beyond the test situation seems difficult if not impossible. On the other hand, high-quality language tests are generally based on something firmer. Using a well-established view of communicative competence such as that described in the CEFR must inevitably lead to greater validity of a test procedure, especially when added to test development procedures such as those described in the *Manual for Relating Language Tests to the Common European Framework of Reference*. A test which is based on descriptors of communicative performance, as described in the CEFR, does not set out to test features of personality and is not based on psychological constructs. This does not mean that such features of personality as tolerance of frustration or open-mindedness, to name but two, are not important in intercultural competence, just that they should not form the basis of a test of intercultural competence.

Every test of language competence represents a compromise governed by rules, for which several possibilities may exist. All of these various possibilities for the realisation of a test construct require, however, that the four elements mentioned above are demonstrated: a widely accepted description of communicative competence, a plausible selection of partial competences, an exact definition of criteria for marking performance and standardised testing and scoring procedures. Our development of the test format is based on the principle that there should be as much authentic communication as possible and as much standardisation as necessary to ensure objective evaluation. The format takes into consideration the testable elements of intercultural communicative competence, including cognitive aspects as well as communicative ability, without entirely ignoring the (in fact not testable) features of personality. The test consists of a written part and an oral part, which evaluate speaking, writing, reading and listening in intercultural encounters. Observable and assessable communicative competence is in the foreground. Theoretical intercultural knowledge is not tested and only awarded importance is as far as it contributes to successful practical intercultural communication.

¹³ Morgeson, F.P.; Campion, M.A.; Dipboye, R.L.; Hollenbeck, J.R.; Murphy, K.; Schmitt, N. Are we Getting Fooled Again? Coming to Terms with Limitations in the Use of Personality Tests for Personnel Selection. In: *Personnel Psychology* 2007, 60, 1029-1049.

For a detailed description of the test format mentioned here, see the test manual as published on www.elc-consult.com

Conclusion

Test developers are accustomed to starting with construct and criteria and conducting the development of a test with these in mind. A team of “language experts” such as ours had no difficulty in accepting the premise that intercultural competence only makes sense if it is understood as intercultural **communicative** competence. It therefore seemed to be the next logical step to find all the relevant statements and descriptors in the CEFR and use these to approach the question of how language competence (as it has been tested for years) and intercultural competence go together and where differences may lie. Using the criterion-based test format and the elements belonging to this – syllabus, material and test as well as train-the-trainer courses – we are convinced that this example of training and testing intercultural competence in English provides a realistic and practicable method which can be transferred to other languages. The performance-based approach to training and testing may also replace the cognitive approach to the training of intercultural competence in use up to now.

References

- Beneke, Jürgen (2000). Intercultural competence. In: Bliesener, Ulrich (ed.) Training the Trainers. Theory and Practice of Foreign Language Teacher Education (2000).
- Bolten, Jürgen (2003). Interkulturelles Coaching, Mediation, Training und Consulting als Aufgaben des Personalmanagements internationaler Unternehmen. In: Bolten, J. / Ehrhardt, C. (ed.), Interkulturelle Kommunikation. Texte und Übungen zum interkulturellen Handeln (2003).
- Byram, Michael (1997). Teaching and Assessing Intercultural Communicative Competence.
- Byram, M. / Nichols, A. / Stevens, D. (eds) (2001): Developing Intercultural Competence in Practice.
- Knoll, Yvonne (2006). Currently Offered Intercultural Training in Germany and Great Britain. An Empirical Study, in: Interculture Journal 2006/1
- Mader, Judith / Camerer, Rudi: International English and the Teaching of Intercultural Communicative Competence. In: Interculture Journal 12/2010. pp. 97-116
- Meierkord, Christiane (1996). Englisch als Medium der interkulturellen Kommunikation: Untersuchungen zum non-native-/non-native-speaker-Diskurs.
- Müller-Jacquier, Bernd (1999), Interkulturelle Kommunikation und Fremdsprachendidaktik. Studienbrief Kulturwissenschaft. (unpublished Koblenz 1999).
- Müller-Jacquier, Bernd (2000). Linguistic Awareness of Cultures. Grundlagen eines Trainingsmoduls. In: Bolten, Jürgen (ed.): Studien zur internationalen Unternehmens-Kommunikation. Waldsteinberg 2000.

Morgeson, F.P./ Campion, M.A./ Dipboye, R.L./ Hollenbeck, J.R./ Murphy, K./ Schmitt; N. Are we Getting Fooled Again? Coming to Terms with Limitations in the Use of Personality Tests for Personnel Selection. In: Personnel Psychology 2007, 60, 1029-1049.

Spencer-Oatey, Helen / Franklin, Peter (2009). Intercultural interaction: a multidisciplinary approach to intercultural communication.

Wolf, Hans-Georg / Polzenhagen, Frank (2006). Intercultural Communication in English – A Cognitive Linguistic Focus on Neglected Issues.

r.camerer@elc-consult.com

j.mader@elc-consult.com

Judith Mader has degrees from the Universities of Sussex and Birmingham. She comes from the North of England and has lived in Germany for over 25 years. She has worked in many areas of ELT, including full-time as a test developer for telc - The European Language Certificates. She is Head of Languages at the Frankfurt School of Finance and Business and also a senior consultant to elc – European Language Competence in Frankfurt. Her research interests are English as an International language, its impact on testing and its influence on intercultural communication.

Rudi Camerer has a PhD in English studies. He lived and worked in the US, Britain and much of Europe for several years. He has directed small and municipal centres of adult education in Germany, including that of the City of Hamburg. He directed the head office of telc – The European Language Certificates before establishing his own consultancy elc – European Language Competence in Frankfurt am Main. His research interests are the effect of politeness discourse conventions on intercultural communication and the use of International English in intercultural encounters.

CEFR and contrastive rhetoric – what's the link?

Cecilie Carlsen,

University of Bergen, Norway



Introduction

An overall aim of the CEFR is to be open, dynamic, non-dogmatic, and general enough to be relevant to different languages, learner-groups and situations (CEFR: 8). To meet this aim the CEFR needs to describe language proficiency in a non language-specific way (Hulstijn, Alderson and Schroonen, 2010:16). This, however, is not unproblematic (Alderson, 2007:660). Firstly, it is likely that the difficulty of reaching a certain level of proficiency in, for instance, grammar, orthography or pronunciation will vary according to characteristics of the *target language*. It is easier to spell correctly if the writing system of the target language reflects pronunciation, which is the case for Spanish and Finnish more than for English and Danish. Similarly, it is easier to master the morphology of English than that of Polish due to the more complex morphological system of the latter; the rules of syntax are more complex in Norwegian than in Spanish and so on. Secondly, the challenges of reaching a certain level of proficiency will vary according to characteristics of the learners' *first language* (L1) and the relative distance between the L1 and the target language. The learners' L1 may affect the time it takes, the degree of success, as well as the very process of learning the target language (Jarvis and Pavlenko, 2007). Empirical research investigating specific first target language combinations in relation to the CEFR is therefore sorely needed, and the network of Second Language Acquisition and Language Testing in Europe (SLATE) deserves to be mentioned in this respect (Bartning, Marting and Vedder, 2010).

In this paper, I will focus on potential L1-transfer at the level of text structure and written discourse in relation to the CEFR. Most of the research on discursive transfer has been carried out within the framework of contrastive rhetoric (Jarvis and Pavlenko, 2007:103) Contrastive rhetoric (CR) considers texts as cultural phenomena, and postulates that norms of text quality vary somewhat from one language community to another (Kaplan, 1966; Connor, 1996; Connor

et al, 2008). A question I find highly relevant is whether the CEFR imposes one particular cultural view of text quality or whether it is generous and general enough to include different cultural norms. During my workshop at the IATEFL/EALTA conference we discussed the following questions:

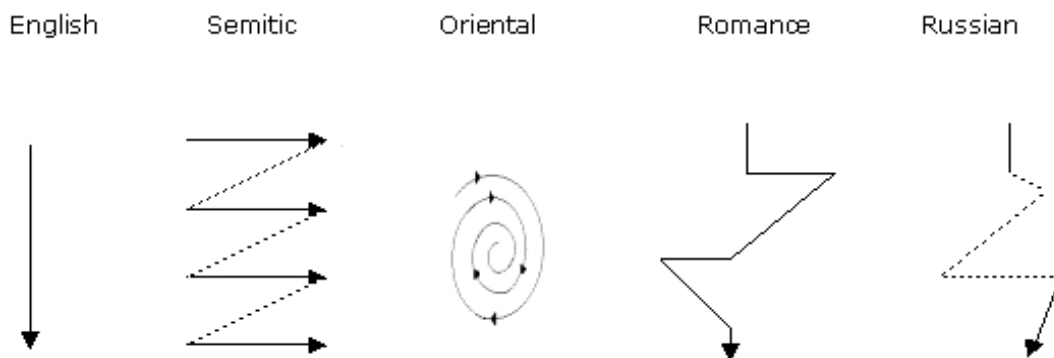
- are there cultural norms for good writing?
- is the CEFR generous and general enough to cover such differences? or
- does the CEFR impose a text norm which is closer to that of some language communities than others?

Results from a small pilot study carried out during the workshop will be presented later in this paper.

Contrastive rhetoric

Contrastive rhetoric has its roots in the 1960s and Robert Kaplan’s seminal article “Cultural thought patterns in inter-cultural education” published in *Language Learning* in 1966. In short, CR predicts that there are differences between languages not only at the level of grammar and vocabulary, but also when it comes to what is regarded as a well-composed text. These different text norms may transfer when learning a new language, resulting in L2 texts which deviate somewhat from the preferred structure in the target language community. The original CR-hypothesis maintains that “[e]ach language and each culture has a paragraph order unique to itself, and [...] part of the learning of a particular language is the mastery of its logical system” (Kaplan, 1966: 14). Based on the analysis of some 600 ESL essays written by students with different L1s, Kaplan identified five dominant paragraph structures, which he illustrated by the frequently-quoted doodles below:

Figure 1: Cross-cultural differences in paragraph organization (Kaplan 1966)



Kaplan’s traditional CR-hypothesis has met with criticism (see Connor, 2002 for an overview),

which has eventually led to a redefinition of the field. Modern approaches to contrastive rhetoric, often referred to as intercultural rhetoric, differ from the traditional approach in many important ways (Connor, 2002; Connor et al 2008): There is a greater striving for methodological rigour, there is more focus on comparing similar texts across languages and cultures, and the concept of culture is redefined in accordance with Holliday and Atkinson's ideas stating that within one and the same national culture, there may be different small-cultures which may differ from other small-cultures within the same national culture (Holliday, 1999; Atkinson, 2004). Finally, Connor and Moreno postulate a new and modified hypothesis of intercultural rhetoric in which they maintain that: "[...] different cultures have different rhetorical tendencies. [...] the linguistic patterns and rhetorical conventions of the first language (L1) often transfer to writing in second language (L2)" (Connor & Moreno, 2005:153)¹⁴.

Contrastive rhetoric and the CEFR

Searches in the electronic version of the CEFR reveal no references to the term "contrastive rhetoric" or "intercultural rhetoric". There are, however, numerous references to differences and similarities between cultures. The CEFR deals with cultural competence both as integrated in the concept of communicative competence, particularly in the treatment of sociolinguistic and pragmatic competence, as well as in relation to the concept of intercultural awareness, defined in the CEFR as the "[k]nowledge, awareness and understanding of the relation (similarities and distinctive differences) between the 'world of origin' and the 'world of the target community'" (CEFR, 2001:103). What seems to be lacking, however, is an explicit discussion of the question whether the CEFR scales impose a particular cultural view of text quality. Whether this is the case should be investigated systematically in relation to the scales relevant for written production.

Workshop: A small pilot study

In the workshop, participants¹⁵ were asked to form L1-groups and discuss what they considered the most important criteria for text quality in their language community. Since this may vary according to text genre and proficiency level, they were asked to think about a typical school essay or argumentative essay written by an L1-user of their language in upper secondary

¹⁴Stutterheim et al (2010) also find that different languages tend to structure texts differently, and that these different tendencies transfer when learning a new language. As opposed to the CR-approach, they argue that these differences are driven by grammatical differences between languages and not by cultural differences.

¹⁵ The participants were not stratified in any way. The only participant profile information I have available is that they are teachers or language testers in different European countries, and since they attended a conference about the CEFR, I assume they have some knowledge about this framework.

school. They were asked to try *not* to think in CEFR-terms but to specify what is normally regarded a good text in their language community¹⁶.

Thereafter, participants were asked to work individually and respond to 16 statements about text quality on a four-point Likert-scale, bearing the same kind of text in mind. The questionnaire contained statements about text norms, in particular those that prior CR-studies had found to vary across cultures, such as preferences for short vs. long sentences, for nominal vs. clausal clauses, for active vs. passive voice, and for elaborate vs. straightforward style (see Connor, 1996 for an overview). In addition it focused on the tolerance for digressions and deviations from the main point, as well as on the use of examples and metaphors. Another important distinction is reader vs. writer responsibility, i.e. whether text comprehensibility is considered mainly the responsibility of the writer or of the reader (Hinds, 1987). Statement 6, 12, 13, and 15 all relate to this distinction.

Results of the pilot study

There were only 25 informants in the present study, and hence the results should be interpreted with caution.

Table 1: Questionnaire about text quality. Results are percentages of 25 participants.

	(a typical argumentative essay/L1-texts/ upper secondary school) A good text...	Agree	Agree some	Dis- agree some	Dis- agree
1	...is clearly and logically organized.	96	4	-	-
2	...sticks to the point without deviations.	60	36	4	-
3	...provides examples to illustrate the main points.	80	12	4	-
4	...uses an elaborate language where one paragraph may well contain only one complex sentence.	4	16	36	36
5	...uses the active rather than the passive voice.	16	32	44	8
6	...gives readers the possibility to use their	32	36	20	8

¹⁶ The results of this initial exercise are not presented here but were important as a consciousness-raising activity.

	intelligence to construct meaning when reading it.				
7	...uses verbal rather than nominal constructions.	16	40	32	12
8	...introduces the main point early.	56	28	12	-
9	...contains deviations and digressions to tickle the readers' interest.	8	24	24	40
10	...is kept in a simple and straightforward style.	52	32	16	-
11	...uses a personal ("I, we") rather than an impersonal style ("one, you, it").	12	28	48	12
12	...makes use of meta-language to tell the reader what to expect (for instance: "In this text, I will..., thereafter..., In the conclusion...")	52	28	16	4
13	...makes sure the reader understands.	68	28	4	-
14	...repeats and reformulates important points throughout.	20	48	20	8
15	...does not spell things out too explicitly.	8	20	48	24
16	...makes use of metaphors.	16	48	28	8
<i>For some of the question the total is less than 100 % because of some missing values. 0 % is replaced by "-" to make the grid more readable.</i>					

For a few of the statements there is a high agreement between the 25 informants. For the first statement "A good text is clearly and logically organized" 96 % of the informants agree. The question is of course whether informants vary as to *what* they consider to be a clearly and logically organized text, which is what the rest of the statements are trying to pin down. "A good text provides examples to illustrate the main points" is supported by 80 % of the informants. 68 % agrees that "A good text makes sure the reader understands". This is an important question as it focuses on reader vs. writer responsibility, as mentioned above. There is less agreement about its counterpart in the grid, which states that: "A good text gives readers the possibility to use their intelligence to construct meaning when reading it". The largest diversion between informants is found for statements 4, 7 and 9.

L1-group differences

The main purpose of the questionnaire was to investigate potential L1-differences regarding criteria for text quality. The 25 participants were divided into five L1- groups: English (n=7), German (n=6), Norwegian (n=6), Finnish (n=3), and Spanish (n=3). L1-group differences were analysed with a One-Way Analysis of Variance (ANOVA) in SPSS. ANOVA is a statistical method that compares the variance *within* groups to the variance *between* groups in order to answer the question whether the differences between groups are sufficiently large to support the claim that the groups represent different populations (Larson-Hall, 2010:268). The results of the ANOVA show that differences between L1-groups are significant ($p > .05$) for 4 of the 16 statements, i.e. statements 2, 7, 8 and 12¹⁷.

Table 2: Results of the ANOVA statistics

Statement	F-values	Sig.
2. Sticks to the point	$F_{4,20} = 4,313$.011
7. Verbal rather than nominal	$F_{4,20} = 2,865$.050
8. Main point early	$F_{4,20} = 3,055$.041
12. Use of meta-language	$F_{4,20} = 4,800$.007

The numbers after the F-value $F_{4,20}$ refers to the degrees of freedom of the independent variable/between groups (4) and the degrees of freedom of the error/ within groups (20).

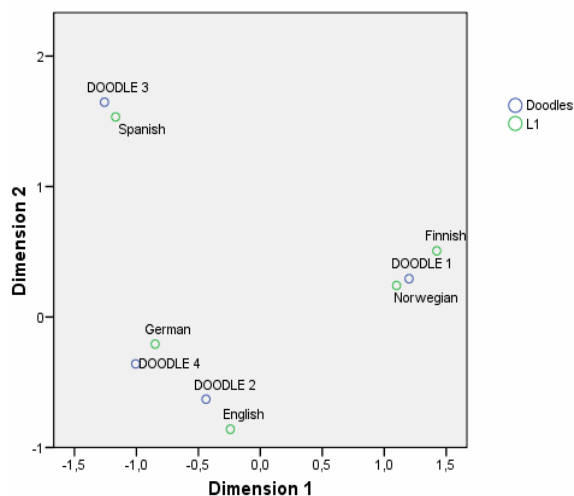
Post-hoc analyses were calculated for those statements that obtained significant F-values in order to answer the question of which L1-groups contribute most to the mean difference between groups for the different statements. Post-hoc analyses using the LSD post-hoc criterion for significance indicate that for statement 1 (“Sticks to the point without deviation”) the German L1 group stands out from the others, differing significantly from the English, Norwegian and Finnish L1-groups ($p > .05$). For statement 7 (“Verbal rather than nominal”), the Norwegian L1-group contributes most to group difference, differing significantly from the English, German, and Spanish L1-groups. For statement 8 (“Introduces the main point early”) again, the German group contributes most, differing significantly from both English and Norwegian ($p > .05$). Finally, for statement 12 (“Use of meta-language”) German differs significantly from English and Norwegian, and the Norwegian and Spanish L1-groups also differ significantly from one another. The results of the post-hoc analysis show that the German L1-group contributes most to the F-values overall, followed by the Norwegian L1-group.

¹⁷ The statements where there were not significant differences between L1-groups will not be further commented on here.

The workshop participants were also asked to mark which of the Kaplan doodles best represented their L1s. The L1-labels of the original doodles were omitted from the informants' questionnaire. L1-group differences for the choice of doodles is significant at the 0.012 level ($p > 0.05$) using the Chi-square test. 6 out of the 7 English L1-informants found that the second doodle from the left (parallel structures) is the one that best represents the text structure of English. 5 out of the 6 Norwegian L1-users on the other hand found that the linear, straightforward structure of doodle 1 best reflects Norwegian writing. Like the English L1-speakers, most of the German informants chose doodle 2 as representative for their language, while one chose the circular nr 3 and one nr 4. Interestingly, no German speaker chose the straightforward, linear structure of doodle 1, which is in line with the results of the questionnaire which shows that, while all of the language groups agree with statement 10 ("A good text is kept in a simple and straightforward style") half of the German group disagreed with this criterion of text quality. It is also interesting to note that while the circular doodle 3 was only chosen by 3 of the 23 informants altogether (2 missing), 2 of the 3 Spanish speakers chose this doodle to best represent the text structure of Spanish. The groups are too small to generalize from the results or to give them much importance, but sufficient to arouse curiosity and inspire further studies.

The differences between the L1-groups' text structure preferences are visualized in the output of the Correspondence Analysis presented below. We see that Finnish and Norwegian have a preference for doodle 1, while Spanish stands out in sharp distinction to the other groups as the only one choosing doodle 3. German and English tend to chose doodles 2 and 4.

Figure 2: Correspondence-analysis between L1-groups and doodles



Discussion and tentative conclusion

The results of this small pilot study yield some support to the findings of numerous studies indicating that the concept of text quality varies somewhat across languages (Connor, 1996; Connor et al, 2008, Stutterheim et al 2010). The results indicate that German differs in many ways from English and Norwegian, which is interesting since these languages often tend to be grouped together under the label of German languages. When it comes to text norms, historical and societal factors play an important role, and even though German and Norwegian share many linguistic commonalities, there are marked societal and historical differences between the two countries that may affect written text norms.

In the workshop we discussed the questions presented in the introduction of this short paper. The main conclusion was that these questions were important and should be given more attention in relation to the implementation of the CEFR across languages and cultures. There was a tentative positive answer to the question of whether the CEFR is general enough to cater for different text norms. However, for this to take place, users need to be aware of the culture-specific norms for text organization when implementing the CEFR in language-specific contexts, for instance when developing language-specific rating criteria for writing, or reference level descriptions (RLD) for different languages. My intention has been to draw attention to these matters, and it is my hope that we will continue to discuss the role of the first language in relation to the implementation of the CEFR in the future.

References

- Alderson, C. (2007) *The CEFR and the need for more research*. In: *The Modern Language Journal*, 91,iv,:659-63.
- Atkinson, D. (2004) *Contrasting rhetorics/contrasting cultures: Why contrastive rhetoric needs a better conceptualization of culture* In: *Journal of English for Academic Purposes*, 3:277-90.
- Bartning, I., Martin, M. & Vedder, I. (Eds.) *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*.
- Connor, U. (1996) *Contrastive Rhetoric. Cross-cultural aspects of second-language writing*. Cambridge: Cambridge University Press.
- Connor, U. (2002) *New directions in contrastive rhetoric*. In: *TESOL Quarterly*, 36/4: 493-510.
- Connor, U. and Moreno, A. (2005) *Tertium comparationis. A vital component in contrastive rhetoric research*. In Bruthiaux, P. et al (Eds.). *Directions in applied linguistics: Essays in honour*

of Robert B. Kaplan. UK: Multilingual Matters.

Connor, U., Nagelhout, E. & Rozycki, W. (2008) *Contrastive Rhetoric. Reaching to intercultural rhetoric*. Amsterdam: John Benjamins Publishing Company

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

Hinds, J. (1987) Reader versus writer responsibility: A new typology. In: Connor, U. and Kaplan, R. (Eds.) *Writing across languages: Analysis of L2-texts*. Reading, MA: Addison-Wesley.

Holliday, A. (1999). Small cultures. In: *Applied Linguistics*, 20: 237-64.

Hulstijn, J., Alderson, C. & Schroonen, R. (2010). Developmental stages in second-language acquisition and levels of second language proficiency. Are there links between them? In Bartning, I., Martin, M. & Vedder, I. (Eds.) *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*.

Jarvis. S. and Pavlenko, A. (2007) *Crosslinguistic influence in language and cognition*. New York: Routledge

Kaplan, R. (1966) Cultural thought patterns in inter-cultural education. *Language Learning: A Journal of Applied Linguistics*, 16:1-20.

Larson-Hall, J. (2010) *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Stutterheim, C.v., Bouhaous, A., Carroll, M., and Sahonenko, N. (2010) Grammaticalized temporal categories, language specificity and macro planning in expository texts. Under review for *Linguistics*.

cecilie.carlsen@lle.uib.no

Cecilie Carlsen holds a Ph.D. in language assessment from 2004. From 2003-05 she was part of a team developing national tests in English for Norwegian school children. Since 2005 she has been working with the assessment of adult immigrants at Norsk språktest. She is currently doing a postdoctoral research project at the University of Bergen, in which she uses an electronic learner corpus to investigate L1-influence on the discourse level.

Putting the CEFR to Good Use – A Collaborative Challenge

Gudrun Erickson

University of Gothenburg, Sweden



Introduction

Since its publication in 2001, the Common European Framework of Reference for Languages, the CEFR (Council of Europe, 2001), has had increasing influence on language policies and practices at local, national and international levels. Implementing this rich and complex document is obviously a challenge in different ways, depending, for example, on pedagogical traditions, contextual factors and political climate. In this keynote paper, collaboration between different stakeholders at different levels of the educational system is emphasized as a powerful means of achieving the goal of Putting the CEFR to Good Use. Examples are given from a national – Swedish – perspective, a joint project between the national and international levels, and, finally, from an international perspective.

Collaboration in a conceptual frame

The conceptual basis for the present paper is a unified and extended view of validity. The traditional definition of this concept implies that you need to make sure that you are actually assessing, or measuring, what should be assessed/measured, nothing else. This means, on the one hand, that you have to check, continuously, that you do not exclude important aspects of what is in focus, usually referred to as the *construct*, on the other hand that you do not include things that really do not belong to this construct. This definition of validity still holds true, but during the last few decades the concept has been extended to focusing on the *use* of the results generated by the assessment. This includes inferences, decisions and actions, and to some extent even consequences, arising from what has been observed (Messick, 1989; Moss et al., 2006; Bachman and Palmer, 2010). Handling this obviously demanding task requires a number of strict principles and procedures. One of these is ensuring that there are multiple sources of evidence underpinning any claims made, another is successive and strict quality control of processes as well as products. In addition, collaboration between different categories of interested parties, often referred to as *stakeholders*, is an important factor in the validation

process, with clear ideological and ethical implications (Shohamy, 2001). Collaboration is also the main theme of the current paper.

An example of collaboration in national test development

There is a long tradition of national testing and assessment in Sweden, aimed at complementing teachers' continuous observations. National as well as local assessments are based on national curricula, which define the goals of schooling in general and those of individual subjects.

Formative as well as summative national assessment and testing materials are provided, both very well received by teachers and students. Different universities, with solid, documented research within relevant domains, are commissioned by the National Agency for Education to take responsibility for the development of the different materials. In the case of foreign languages – English, French, German and Spanish – this is done at the University of Gothenburg.

The development of the national testing and assessment materials of foreign languages is based on publicly available principles, common for all materials, and on specifications for the different tests (http://www.nafs.gu.se/english/information/nafs_eng/). The development process is distinctly collaborative, with systematic involvement of different categories of stakeholders, all of them contributing their own special expertise. The most important partners in the process are teachers, teacher educators, researchers from different disciplines, and, perhaps most importantly, students of different ages. There are several reasons why test-takers are essential in test development; Motives related to ethics and democracy can be brought forward, as can aspects of pedagogy, impact, and – obviously – validity in a wide sense, where the use of results and the effects of the inferences made are in focus. The most important reason to actively collaborate with students, though, has to do with what is perceived as sheer necessity. However well educated and experienced, test developers can never fully foresee the interpretations and reactions of a wide group of test-takers, a number of them probably not quite as enthusiastic about language and language study as those in charge of producing the tests. Thus, students' input provides unique information, essential for the quality of the testing materials.

The development process comprises successive meetings with reference groups, whose members, to a varying extent, take part in discussions, item writing, analyses of results, decisions about test composition, and standard setting. There are consequently no item writers

in the isolated sense; people always have more than one function in the development process. This is believed to contribute to quality in a wide sense, which can be seen as confirmed by reliability and validity analyses with satisfactory results, and by the high degree of acceptance of the national tests among teachers and students. All materials are piloted successively and, after different modifications, pre-tested in large, randomly chosen samples in the country (usually around 400 students per task). In connection with this, all teachers and students answer extensive questionnaires, the results of which are actively used in the analyses of the results. Aspects commented on by both teachers and students, although obviously phrased somewhat differently, concern, for example, perceived relevance, difficulty and clarity of different tasks, and also the degree of familiarity, i.e. how commonly used different tasks are in the regular teaching and learning situation. In addition, students are asked to make retrospective, task-related assessments of their own performances. The questionnaires comprise multiple-choice questions for teachers and Likert scales for students. For both categories there is also ample space for personal comments.

The contributions of the different stakeholder groups are diverse. Teachers and teacher educators are active members of reference groups, where researchers also come in, often in connection with special analyses and studies. Examples of the latter comprise different aspects of language and language use, analyses of inter-rater consistency, studies of the dimensionality of language performance, and different analyses of rater behaviour, for example involving reflective protocols. Students contribute in many ways, e.g. by providing information that helps to enhance the quality of tasks as well as teachers' guidelines. Their input also affects the selection of topics and tasks, thereby, hopefully, giving test-takers optimal chances to show what they actually *know* and *can do* with their language. Last but not least, student feedback helps in the sequencing of tasks within a test, something that obviously affects overall performances. (For additional information, see Erickson, 2009; Erickson, 2010).

A national and international example of collaboration

For a long time, the Swedish national syllabuses for foreign languages have had a distinctly functional and communicative character, very similar to what in the CEFR is referred to as "an action oriented approach". This is reasonably well implemented in Swedish language classrooms. However, the levels of proficiency required for the different stages in the system, defined by national goals and grading criteria, have not yet been fully aligned to the CEFR levels, and scaling as such is a novelty to quite a number of teachers. Some tentative textual

analyses have been made, as well as continuous, empirical observations related to national test development, but so far no large-scale, systematic studies have been performed. However, in connection with the current revision of the syllabuses, there is a clearly stated ambition to bring the Swedish system nearer to the CEFR, this time also including alignment of proficiency levels. Initiated by the national test development group at the University of Gothenburg, and funded by the Swedish National Agency for Education, a tentative, initial study was therefore designed, aimed at investigating one of the already existing national tests in relation to the CEFR.

Twelve experts, all with profound professional experience of the CEFR and of language testing and assessment, in twelve different countries, generously consented to participate in the study. The aim was to tentatively analyse and relate the national test of English for the end of compulsory school in the Swedish system to the CEFR, with regard to tasks as well as to standards. The study, undertaken in the late spring of 2009, followed a basic scheme, however with ample opportunity for the participants to comment freely on any aspect they found relevant. With the aim of receiving independent judgements, the participants were not informed about each other.

The informants were given full background information about the test, through translated documents, teacher guidelines, including benchmarks and cut-scores, and an article about relevant framework factors and national test development (Erickson, 2009). They were also provided with the actual testing materials, which, like all other national tests of foreign languages in Sweden, are monolingual, i.e. using the target language only.

Self-evidently, the length and wealth of details in the informants' reports varied, however with excellent overall quality. Many aspects of interest, both concerning test development, interpretations and applications of the CEFR, and the Swedish system of testing and grading at large, were highlighted. A number of positive comments were given, but also some doubts were expressed about the ability of one test to cater for the obvious heterogeneity of whole cohorts, which is the case in the Swedish school system. Examples of other issues that demonstrated a certain variability of opinions concerned the degree of standardization and the content and format of individual tasks. As for the main research issue, the relation between the test and the CEFR, the overall result confirmed the opinion commonly held in Sweden, namely that the Pass level of the test is reasonably equivalent to a low B1. As for the highest grade level, a Pass with special distinction, the informants generally thought that the students at the end of compulsory

school belonging to this category – roughly about 16 per cent – demonstrate language proficiency at a (high) B2 level.

As pointed out initially, the project reported here was indeed tentative; thus, no far-reaching conclusions should be drawn on the basis of the results. Rather, it should be seen as providing interesting indications and collegial reflections of great value to the continued Swedish efforts to align the national syllabuses to the CEFR. Furthermore, it is a good example of the mutual value of collaboration, where all parties involved gain insights beneficial to the efforts of further enhancing the quality of processes as well as products within the field of language testing and assessment.

International examples of collaboration in language testing and assessment

There are numerous examples of successful collaborative projects within the field of language learning, teaching and assessment in and outside of Europe. During the last decade, many of these have been related, one way or the other, to the CEFR. Initiatives have been taken by the Council of Europe, the European Centre for Modern Languages, the European Commission, individual countries, universities and other institutions, companies, *and*, not to be forgotten, associations and groups like the IATEFL TEASIG and EALTA. The very conference where this plenary paper was presented is a good example of such an initiative, aimed to enhance the good use of the CEFR.

Of the many successful endeavours undertaken, it seems relevant to refer to the ENLTA project – European Network for Language Testing and Assessment – run between 2003 and 2005, funded by the EU, and aimed at supporting the initial development of the European Association for Language Testing and Assessment (EALTA). The project, coordinated by Professor J. Charles Alderson at Lancaster University, UK, defined eight different activities, all with their own project leaders. The studies undertaken were reported successively, and some of them are currently available on what has later become the EALTA Resources webpage (www.ealta.eu.org/resources.htm). One of the published studies explored Language Testing and Assessment Needs in Europe, the first part of the survey giving a general picture (Hasselgreen et al., 2004), the second reporting on regional needs (Huhta et al., 2005). The findings were quite unanimous, namely that all three groups approached, viz. teachers, teacher educators and people involved in large scale test development, considered themselves in need of more training in different types of assessment, comprising both so-called alternative methods and more

psychometrically oriented techniques. Another study focused on views of language testing and assessment among language teachers and learners (Erickson & Gustafsson, 2005). Information was collected through questionnaires, with responses obtained from 1373 teenage students and their 62 teachers in ten European countries. One important outcome of the survey was the substantial input given by the students, who willingly and reflectively shared their experiences and opinions, another the clear similarities between the views of teachers and students. Furthermore, the results demonstrated clear correspondence with what has been noted over the years in the collection of test-taker feedback in connection with national test development in Sweden.

Finally, an obvious example of collaboration between a large number of people, institutions and countries is EALTA as such, officially founded in 2004, with the ambition of reaching a wide membership including teacher educators, teachers and large-scale test developers, and with a Mission Statement making explicit the value of collaboration:

The purpose of EALTA is to promote the understanding of theoretical principles of language testing and assessment, and the improvement and sharing of testing and assessment practices throughout Europe (<http://www.ealta.eu.org>)

There are many manifestations of this aim, for example that

- a discussion list for members is provided
- there are Special Interest Groups focusing on different aspects of language testing and assessment
- all materials published on the Resources website (<http://www.ealta.eu.org/resources.htm>) can be downloaded free of charge
- there is a variety activities offered for members
- costs, both for administration and for participation in the events offered, are kept at an absolute minimum, reflecting the inclusive aim of the association.

Perhaps the most obvious manifestation of the Mission Statement, and of the fruitfulness of collaboration, is the development of the Guidelines for Good Practice in Language Testing and Assessment, initially drafted within the ENLTA project, further developed by a special working group, discussed among the members, adopted in 2006 and currently available in 35 language versions. In these guidelines, basic principles for all types of assessment are emphasized, but

there are also sections for each of the three membership categories, with questions about principles and procedures (<http://www.ealta.eu.org/guidelines.htm>). The document is used in many different contexts in and outside Europe, in pre- and in-service teacher education and in research of various kinds, one example given in Alderson (2008). The Guidelines have also been disseminated in large-scale regional and national events (Erickson & Figueras, 2010).

Concluding reflections

The Common European Framework of Reference for Languages is a rich and complex document with a strong impact on learning, teaching and assessment, in and outside Europe. Furthermore, the CEFR affects national and international policy decisions and actions, thus reaching beyond the important pedagogical uses often discussed. In order to facilitate and optimize the interpretation and implementation of the CEFR, collaboration between users at different levels is called for. Indeed, collaboration as such is a challenge, requiring readiness and ability to broaden perspectives, to view things from other angles and, when proven warranted, to modify traditional procedures and products. As hopefully shown through the examples given in the current paper, however, such endeavours are certainly worthwhile to achieve the goal of Putting the CEFR to Good Use.

References

- Alderson, J. C. (2008) *Final Report on Aviation English Testing*. Retrieved on 14th December 2010 from: <http://www.ealta.eu.org/guidelines.htm>
- Bachman, L. & Palmer, A. (2010) *Language Assessment in Practice*. Oxford: Oxford University Press.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Erickson, G. (2009) *National assessment of foreign languages in Sweden*. Retrieved on 14th December 2010 from http://www.nafs.gu.se/english/information/nafs_eng/
- Erickson, G. (2010) Good Practice in Language Testing and Assessment – A Matter of Responsibility and Respect. In Kao, T. and Lin Y. (eds.), *A New Look at Teaching and Testing: English as Subject and Vehicle* (pp. 237-258). Selected papers from the 2009 LTTC International Conference on English Language Teaching and Testing. March 6-7, Taipei. Taiwan: Bookman Books Ltd (ISBN 978 -957-28764-2-8).
- Erickson, G. & Figueras, N. (2010) *EALTA Guidelines for Good Practice in Language Testing and Assessment: Large scale dissemination days*. Retrieved on 14th December, 2010 from:

<http://www.ealta.eu.org/guidelines.htm>

Erickson, G. & Gustafsson, J-E. (2005) *Some European Students' and Teachers' Views on Language Testing and Assessment. A report on a questionnaire survey*. Retrieved on 14th December, 2010 from: <http://www.ealta.eu.org/resources.htm>

Hasselgreen, A., Carlsen, C. & Helness, H. (2004) European Survey of Language Testing and Assessment Needs. **Report: part one – general findings**. Retrieved on 14th December, 2010 from: <http://www.ealta.eu.org/resources.htm>

Huhta, A., Hirvelä, T. & Banerjee, J. (2005) European Survey of Language Testing and Assessment Needs. **Report: part two - regional findings**. . Retrieved on 14th December, 2010 from: <http://www.ealta.eu.org/resources.htm>

Messick, S. A. (1989). Validity. I R. L. Linn (ed.) *Educational Measurement* (Third edition, pp. 13-103). New York: American Council on Education/Macmillan.

Moss, P., Girard, B. & Haniford, L. (2006) Validity in Educational Assessment. In *Review of Research in Education*, 30(1), 109-162. London: Sage Publications.

Shohamy, E. (2001) *The Power of Tests: a critical perspective on the uses of language tests*. Harlow: Pearson Education.

gudrun.erickson@ped.gu.se

Gudrun Erickson is Associate Professor at the University of Gothenburg, Sweden, and Secretary of EALTA. Originally a language teacher with long experience of teacher education. Project leader for the Swedish national testing and assessment programme for foreign languages, with experiences from a number of European projects. Main research interest in collaborative approaches to test development, in particular contributions by test-takers.

A Virtual Approach to CEFR at University Levels

Isabel Herrando Rodrigo (University of Zaragoza. Spain)



Abstract

The *EEES* is nowadays an opportunity to introduce indispensable traces of the Common European Framework of Reference for Languages (CEFR) at Spanish universities. However, this constant change forces teachers to introduce classroom activities which should be familiar to our 21st century students.

In this project I have aimed to combine new technologies and an introduction to the linguistic competence improvement by means of a virtual *Portfolio*. Students were encouraged to write compositions that were sent to me, corrected, commented in class and sent them back to be incorporated in the students' files or virtual *Portfolios*. By way of conclusion, results show that the use of new technologies boosts communication among students and teachers, improves our tutorial coherency and contributes to putting the CEFR to good use.

Introduction

When we teachers in higher education think about the new items and variables that have to be taken into account at the present moment, we sometimes shake in panic. New ways of teaching are demanded by our students who have been brought up in the *New Technologies Era*. Now that we have to adapt the much needed prevailing knowledge to the new European Framework of Higher Education, new changes could be conceived as potential threats. We may feel that we have many disadvantages against our students because we might think that we “under-use” these new technologies. If we feel so overwhelmed, we can hardly decide if we prefer “trick or treat” or just want to run away from our offices and faculties. With this piece of research I aim to show that what has been possibly seen as a “social threat” around the world can also be used as a teaching tool: *MSN MESSENGER*.

We cannot ignore the fact that we live in a *Society of Information and Communication*. With Wise (2005), Lara (2004) claims that “web-tools” available nowadays boost the learning-teaching process under the umbrella of constructivist pedagogy. Besides, our students are more and more exposed to the virtual Web 2.0 concept of the Internet that forces users to collaborate actively in these new spaces. Edublogs and Wiki-spaces are just two examples of the wide panorama.

It is widely known that our students have expanded their social networks by the use of websites such as Facebook, Twenty, Twitter or Messenger. I therefore took advantage of a very cheap and easy internet tool –*MSN Messenger* – to enable students to communicate with teachers in a practical, efficient and instant way and improve the quality of tutorials by holding “teletutorials”. Above all, this means of communication aims to contribute to the improvement of current practice and change management and to competence building using the CEFR, not only in exam contexts but also in classroom assessment through the use of *Messenger*. Thus, I will show how this channel of communication enhances the constant use of virtual *Portfolios* in the teaching-learning process of my university students. This project is framed in a wider one supported by my *Vicerrectorado de Investigación* from the University of Zaragoza as a part of an *Innovative Teaching Project* (PESUZ 09-5-6).

How can we use the CEFR to improve current practice and manage change?

When thinking about the CEFR levels, we may think about Cambridge ESOL exams for instance:

CPE (Proficiency): C2	MASTERY
CAE (Advanced): C1	EFFECTIVE OPERATIONAL PROFICIENCY
FCE (First): B2	VANTAGE

ILEC or TKT among others.

Our students are expected to enter university with a *Threshold* level of English (B1) from secondary education. With the present project I aim to expand this band towards B2 (Vantage) by working on the Internet sending tasks and feedback to students in order to enable them to create virtual *Portfolios*. Thanks to this system, reflections about language and tasks can be stored for life.

Thus, 2 out of the 10 points from the final mark were evaluated from the tasks carried out with the students' virtual *Portfolio*.

Every week, students were sent different tasks that had to be fulfilled and sent back to me to be assessed. They received my feedback and each week I projected tasks from different students in order to reflect together about the weak and strong points of the task. When a personal or common reflection was useful to the entire group I could send it to all my students with just a mouse click.

Turning our attention to the linguistic features of this project, a special emphasis should be paid on how CEFR can be applied to the teaching of English as ESP in the different degree courses. This may also be formalized in terms of *competences* in the CEFR. When we think about linguistic competence we may remember Chomsky (1979) and his emphasis on the capacity of creating and transforming reality around us by means of language. The CEFR reinforces the idea of combining all the competences. We are “competent” when we apply different knowledge to different situations and with different purposes. Among other researchers, Caturla (2008) claims that we have to be competent citizens and thus, we have to use competences in every context of life in real and relevant scenarios.

I therefore conceive the basic competences as multifunctional everlasting packets that can be used all our lives. These competences should integrate knowledge, procedures and attitudes. Together with this, EEES also highlights the dynamic dimensions of our students under the *Bologna* terms of higher education. In this project I aimed to make students linguistically competent under the umbrella of the CEFR. By *linguistically competent* I understand that students should express themselves in a coherent way, whether in writing or speaking. They should interpret, listen, read and comprehend oral and written messages. Besides this, they should manage their thoughts, emotions, experiences and opinions in order to express their communicative intentions correctly. Hence, as I will show in the following sections that it may be claimed that this project has been a useful tool to put the CEFR to good use by the use of *Messenger* and *Virtual Portfolios*.

Method

The use of MSN Messenger aimed to encourage students’ learning by the exchange and sharing of different activities and by interaction among students and teacher. The e-mail address of the English subjects was CienciasSaludIngles@hotmail.com and the subjects involved were *English for Nurses*, *Technical English for Physiotherapists*, and *English for Occupational Therapists*.

MSN MESSENGER: Creating Virtual Portfolios

Students were asked to use their *Hotmail* address. If they did not have one, I asked them to create one. I explained that this project was framed in a research project and their addresses would always be preserved as confidential. Then, I displayed my *Hotmail* addresses and created a group of favourites with the students’ addresses.

Through *Messenger* (real-timed conversations or email) students were able to communicate with the teacher fluently. Students could contact me easily and could also establish one-to-one meetings. Moreover, those students, who could not attend class, could follow up the subject without stress using this system. As it has been previously commented, different tasks were sent, evaluated and shared with all the students by means of *Messenger*. Thanks to the contact list, files, pieces of advice and different sorts of information were distributed to all the students in seconds.

Evaluation

Student's feedback was collected by means of anonymous questionnaires. The quantitative and qualitative analysis was obtained from the students' data regarding the percentage of students' participation, their satisfaction with their *Portfolios*, the usage of the real-timed conversations, the use of email as a mean of communication or as a means for establishing one-to-one or group seminars, etc. The data was measured using PSP to gain statistics

As I have previously commented in the second section of this article, in the subjects I was in charge of, two points out of 10 were evaluated through compositions sent to the teacher by Messenger. This means of communication made the creation of the *virtual Portfolio* possible. Moreover, getting a mark from this activity made students more willing to collaborate with the project because a part of their final mark depended on it.

Results

Compositions and tasks were projected in class and the corrections were sent back to the students. Students created then *virtual Portfolios* to keep their compositions and tasks. They could therefore reflect on their writing processes. One of the threats of this academic year was the tentative threat of students' failure due to the necessity of finishing their diploma and their massive number of registered subjects. Thus, *Messenger* guaranteed keeping in contact with them. Graph 1 summarises the students' opinion regarding the use of MSN *Messenger* to exchange essays and tasks. From the different options for assessment of this innovative teaching project, "adequate" and "very adequate" were the two options chosen in all the questionnaires. There were no students who claimed that working with *Messenger* seemed "not adequate" or "a failing initiative".

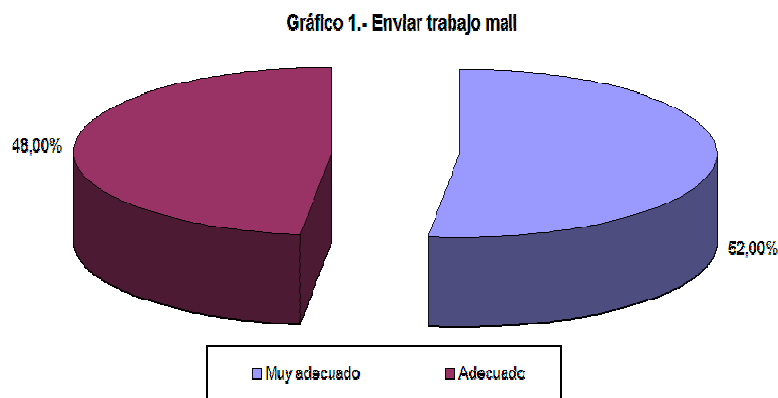


Fig. 1. Students' opinion about the use of *Messenger* as a teaching tool.

As it can be observed from fig. 1, 48% of the students considered this project —using *MSN* for creating a *virtual Portfolio*— useful for reflecting on their learning process. The other 52 % considered it extremely adequate.

Fig. 2 shows the different marks students gave. 64% of the students gave 10 out of 10.

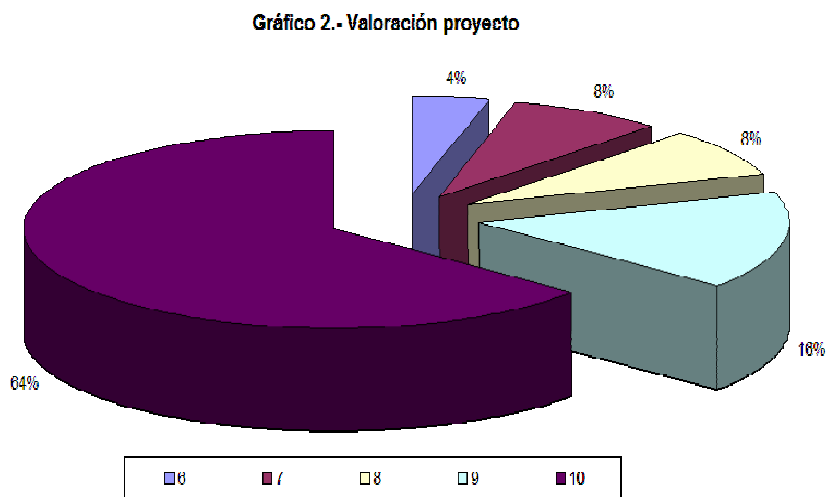


Fig. 2. Students' overall marking from 1 to 10.

96.15% of the students who could not attend class commented on their questionnaires that this project (*MSN* + *virtual Portfolio*) has enabled them to keep the rhythm established in class and learn at the same time. 98.7% of the students who regularly attended class considered that this project contributed positively to the general follow-up of the subject.

Conclusions

Messenger enables the establishment of a virtual space for the exchange of ideas. The use of

Messenger for pedagogical purposes also boosts the communication between students and teachers. It also increases tutorial coventness by creating common interest meeting spaces. This tool reinforces students' confidence through the accessibility of teachers and furthermore the accessibility of the solutions of weak points or problems with the subject.

As a whole it could be affirmed — supported by these results — that *Messenger* (real-timed conversations and email to exchange exercises and doubts) is an innovative tool that encourages a dynamic improvement in current practice and change management despite the tentative social threat of its “agentless” or “de-humanising” characteristics as far as human social relationships are concerned in other spheres of everyday life. The results indicate that almost 97% of the students were aware of their learning process. Thanks to the virtual *Portfolio* updated and “fed” by the use of *Messenger* students admitted to being conscious of their own improvements autonomously. By way of final conclusion it may be claimed that *New Technologies* could be a very useful tool in order to put the CEFR to good use, reinforcing at the same time the linguistic competences suggested by the European Space of Education.

References

- Caturla, E. (2008) “¿Qué hacer con las Competencias?” *Organización y Gestión de Educativa*. Vol 5. 2008.
- Chomsky, N. (1979) *Reflections on Language*. Barcelona: Ariel.
- Dickey, M. D. (2005) “The impact of web-logs (*blogs*) on student perceptions of isolation and alienation in a web-based distance-learning environment”, *Open Learning*, Vol. 19 (3) noviembre 2004.
- Lara, T. (2004) “Nuestros blogs”, *Ciberperiodismo*
<http://blogs.ya.com/ciberperiodismo/>, 19/12/2004.
- Wise, L. (2005) “*Blogs* versus discussion forums in postgraduate on line continuing medical education”, *BlogTalk conference paper*, Sydney, 2005.
http://incsub.org/blogtalk/?page_id=106

herrando@unizar.es

M^a Isabel Herrando teaches Language and Linguistics at the Faculty of Medicine and Arts and English Faculty, Zaragoza, Spain, and holds a MA in Textual and Cultural Studies in English. Her research interests include the study of scientific discourse and language and translation. Her published papers have appeared in *Cambridge University Press*, *Cambridge Scholars*, *Miscelánea*, *Tropelías*, *Prensa Universitaria Zaragoza* and *Journal of the English for Specific Purposes Special Interest Group*.

A Critical Look at the CEFR “Phonological Control” Grid

David Horner PhD

ENSAE ParisTech



Neither the *Common Reference Levels: global scale* nor the *self-assessment grid* makes any reference to pronunciation, which is also conspicuously missing from the subcomponents of the speaking skill in the “qualitative aspects of spoken language use”. So, either the CEFR believes that pronunciation is not important; or that there is sufficient consensus for pronunciation not to need to be emphasized. I examine each of these propositions in turn, before examining critically the single grid for Phonological Control which the CEFR proposes.

1. Is Pronunciation Important?

Of the 30-odd grids covering speaking, only one refers to pronunciation. Yet how one sounds impacts profoundly on how one is perceived. Hieke A.E. (1984), for instance, found that phonological features are an essential element in decoding fluent speech and Mey J. (1998) that native speaker listeners are more intolerant of pronunciation errors than lexical or syntactic errors. There is substantial evidence that non-native accents are subject to negative evaluations by native speakers, and may “be personally downgraded because of their foreign accent” (Leather 1999: 35) and accorded “a lack of competence in many spheres” (Ryan 1983: 155).

Other factors come into play and interact here, notably: (1) the “degree of accentedness” (Ryan & Carranza 1976, Sebastian et al. 1978); (2) the interaction of this element with speakers’ speech styles and social class background (Ryan 1983: 154); (3) the status that some foreign accents have for certain groups of native (and non-native) speakers. In the US, for instance, a Spanish accent in English is more prone to stigmatization than a German one (Bresnahan et al. 2002). Delamare (1996) found that American listeners viewed speakers with certain foreign accents, such as Arabic and Farsi more favourably if the individuals made grammatical errors than if they did not, whereas speakers with other accents, such as French and Malay were actually downgraded if they produced such errors. This implies that the social context in which

native speakers encounter a foreign accent plays an important part in their evaluation of the accent concerned. Moreover, what native speakers will accommodate in a friend they may object to in a professional exchange.

Of course, not all foreign language interactions involve native-speakers (NS), so it is important to determine whether and to what extent tolerance to foreign accent is greater among non-native speakers (NNS), and whether poor pronunciation represents a barrier to understanding.

Jenkins (2000) proposes that many NNS lack the language proficiency to handle speech other than through bottom-up processing, whereas NSs are essentially top-down processors. Although this is probably an oversimplification, it does suggest that NNSs will be more reliant on segmental rather than suprasegmental features; ie the very aspects of pronunciation which a strong accent typically distorts. This is borne out by research indicating that NSs and NNSs do not grade NNS pronunciation according to the same criteria: the former tend to react to suprasegmentals, whereas the latter are more sensitive to segmental features (Johansson 1978, Anderson-Hsieh et al 1992, van den Doel 2006). This may go some way to explaining why Dutch high school students' judgments of the overall proficiency of NN Dutch speakers of English depend more closely on how good their accents were thought to be (Meijer, 2010). In addition, there is evidence that NNSs find foreign accents harder to deal with when faced with faster speech (Rogerson Revell 2010; Anderson-Hsieh and Koehler, 1988). If they are genuinely bottom-up processors, this is not surprising.

Counter-intuitively, perhaps, there is evidence that NNS judges and instructors evaluate foreign learners' errors considerably more severely than do NS and non-instructors (Koster & Koet (1993), Hughes & Lascaratou (1982), Sheorey (1986), Galloway (1980), Schairer (1992), Fayer & Krasinski (1987)).

2. Intelligibility

If one accepts that pronunciation is, at worst, an important element in oral proficiency, then, the degree of intelligibility of the person's speech seems to have become the yardstick by which to measure it. Abercrombie (1956: 93) described "comfortably' intelligible [as] a pronunciation which can be understood with little or no conscious effort on the part of the listener". However, intelligibility is an issue for English in three respects: as an international language, English is characterized perhaps more than any other by its varieties. Moreover, a study by van den Doel

(2006) found that NSs with different varieties of L1 English reacted differently to NNS pronunciation errors.

Because dialects demonstrate linguistically consistent shifts from whatever the local norm is accepted to be, speakers of alternative dialects can usually accommodate fairly rapidly when faced with such accents. However, the situation becomes more complicated with non-native accents because it is significantly easier to understand an accent you are used to than one that you are not used to.

English is not spoken only by and between native English speakers. Its role as an international lingua franca means it is frequently used between native and non-native speakers, or wholly between non-native speakers. The requirement for intelligibility becomes thus more important while remaining equally problematic. On the other hand, speakers who engage often in such exchanges will also gather considerable exposure to different varieties of English.

And this is the crux of the matter: NNSs of English are usually good at deciphering the English they are used to hearing, and much less so when faced with “non-standard” varieties (ie when they move from recorded course book dialogues to listening to real native speakers). In short, the intelligibility issue is not just one of pronunciation: it is also one of listening comprehension. (For a recent discussion of the different factors which adversely affect non-native as opposed to native listening, see Cutler et al. (2004).)

That intelligibility is not just a case of accent is borne out by Munro & Derwing (1995) who found that strongly accented speech cannot be equated with a lack of intelligibility. Similarly, Johansson (1978: 6) points out that speech can be severely distorted and yet be intelligible “... [t]o be communicatively effective, the message must get across swiftly and unambiguously and *without undue demands upon the receiver*” (my italics).

This brings us back to Abercrombie’s definition. However, one person’s perception of effort is not the same as another’s. Our reaction to the effort required to understand someone is related to our experience of hearing speakers of other languages speak ours and to our inherent patience. Moreover, we need to make an effort to understand someone for several interrelated reasons (their poor control of grammar, limited lexis, tiredness ...) which inevitably interact with both a strong accent and the discourse context.

Interestingly, a study by Piazza (1980, 424) into the reactions of French secondary school pupils to grammatical mistakes made by American learners of French found that “[i]rritation was judged more severely than lack of comprehensibility”, especially in spoken language. This is supported by findings from van den Doel’s mammoth 2006 study of NS reactions to Dutch pronunciation. He found that “intelligibility is not the sole criterion used by native speakers in deciding whether a particular pronunciation error is acceptable. *Respondents’ emotive reactions to certain stigmatised realisations indicate that factors such as irritation or amusement also play a part in prioritising certain errors over others.*” Moreover “Respondents’ comments indicate that, across different groups of native speakers, some errors are clearly and consistently more irritating than others (p287).”

Significantly, this finding could be taken to indicate that foreign accents are not only judged on the basis of intelligibility, but also by L1 standards for acceptability, at least where NSs are involved.

So just what is it about accents that renders them more or less difficult to understand or more or less irritating?

3. Is there a Consensus?

What constitutes “acceptable” pronunciation? Is it a fixed variable or one that varies with level? Indeed, what constitutes “pronunciation? And to what extent can one expect learners to master these different elements at different periods in their learning? Or should pronunciation be seen as a holistic variable? Should we therefore be looking at pronunciation in terms of some global marker of intelligibility? In terms of testing, these questions are crucial:

3.1 What is the construct underlying pronunciation?

Can and should learners at different levels of proficiency be assessed on the same elements of this construct? In other words: is there a hierarchy of learning difficulty? Is pronunciation better assessed holistically or analytically?

3.2 Accent: from construct to assessment grid

The construct is the underlying description of a competence that enables a tester to write descriptors that will allow marking grids to be developed. There is general agreement as to what

the different factors are that together make up phonological competence; the CEFR, for instance, lists skill in the perception and production of:

- the sound-units (*phonemes*) of the language and their realization in particular contexts;
- the phonetic features which distinguish phonemes (distinctive features);
- the phonetic composition of words;
- prosody;
- sentence stress and rhythm
- intonation;
- features of linking.

But is each feature is relevant at each level? The CEFR states that “users of the Framework may wish to consider and where appropriate state:

- what new phonological skills are required of the learner;
- what is the relative importance of sounds and prosody;
- whether phonetic accuracy and fluency are an early learning objective or developed as a longer term objective” (p 51.)

Van den Doel’s (2006) study is of great interest here as he found that:

- errors involving word stress were considered to be among the most important;
- much less significance was accorded to the avoidance of weak and contracted forms;
- intonation errors were rated among the least important;
- there was a general tendency for phonemic errors to be ranked more highly than sub-phonemic.

And this, in turn, can be linked to the notions of bottom up and top down processing mentioned earlier. If less proficient learners are more heavily reliant on bottom up processing – related to segmental and word stress features – then it is control of these, I would suggest, that we should be assessing at the lower levels of the CEFR, with the suprasegmentals kicking in later, say as of B2.

3.3 This is where the criticism begins

If the CEFR states that the appropriate criteria are phonetic accuracy and prosody, then one would expect to find them occurring in the grid for phonological control. Yet what we find there are references to concepts like foreign accent, intelligibility, clearness and naturalness:

	PHONOLOGICAL CONTROL
C2	As C1.
C1	Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.
B2	Has acquired a clear, natural, pronunciation and intonation.
B1	Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.
A2	Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.
A1	Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.

Here we see clearly an overt hierarchy of phonological components which learners are expected to master as they progress in proficiency, a hierarchy which we can summarize as follows:

C2	The ability to use intonation and prominence correctly and to express fine nuances of meaning.
C1	The ability to use intonation and prominence correctly and to express fine nuances of meaning.
B2	Clear and natural use of intonation, stress, and sounds.
B1	Intelligible, although a foreign accent is sometimes detectible. Pronunciation errors may occasionally occur.
A2	Pronunciation is sufficiently intelligible, despite a strong accent. The speaker may sometimes have to repeat what they said.
A1	Intelligible with effort.

However, I would take issue with this in several respects. The grid seems to have got the place of intonation right, leaving it until the upper ranges. However, Van den Doel found that the feature identified as most important was word stress. Yet this only appears – perhaps, since the term can just as easily cover prominence – under “stress” at B2. There is no other mention unless it is also subsumed under “pronunciation errors”, in which case the term is far too vague. As it stands, the CEFR suggests that accent disappears beyond B1. Yet “every speaker of a language necessarily speaks it with some accent or other” (Trask 1996: 4).

Intonation is a notoriously difficult skill to acquire, and to suggest that learners at C1 should already master it so as to express fine nuances of meaning seems to me highly ambitious. I would suggest rather that this is a skill which distinguishes the C2 learner. Moreover, as we have seen, the research indicates that it is less relevant to NNS and that even NS find other aspects of pronunciation to be more important. It is therefore right to find it in the upper levels. I cannot

see anything that could distinguish between “clear” and “intelligible”. Proposing that these terms can distinguish between levels therefore appears to me to be unconvincing.

As we have already argued, intelligibility is a much less stable indicator than we would like it to be: much as I feel intuitively that it should be an essential aspect in assessing a learner’s pronunciation, it has to be admitted that we do not all have equal tolerance to foreign accents. This does not, of course, mean that it should not appear as a criterion in assessing pronunciation. But it does underline the necessity of training oral examiners to become accustomed to other accents and to overcome their prejudices. Can one seriously propose that a learner can progress beyond a pronunciation that is “clear and natural”?

I would therefore suggest something like the following simplified version, with the added assumption that proficiency at the level stated indicates that the learner is at the top end of the level, and that most learners assessed will be somewhere between two descriptors (or even a mix of several).

C2	Intonation can be used to express finer shades of meaning. Prominence can be used to express finer shades of meaning
C1	Prominence is used to effect. Features of linking are common Intonation can be used to effect
B2	Sounds and word stress are clearly intelligible. Features of linking appear. Prominence used to effect. Basic intonation patterns are common.
B1	Sounds and word stress are intelligible. Prominence sometimes used to effect Basic intonation patterns appear
A2	Sufficient command of sounds to be understood most of the time. Sufficient command of word stress to be understood most of the time.
A1	Interlocutor will need to ask for repetition and clarification

References

- Abercrombie, D. (1956) “Teaching pronunciation”. In D. Abercrombie, *Problems and Principles: Studies in the Teaching of English as a Second Language*. London: Longmans. 28–40.
- Anderson-Hsieh J. and Koehler, K. (1988) “The effect of foreign accent and speaking rate on native speaker comprehension” *Language Learning* Vol. 38/4: 561–613.
- Anderson-Hsieh J., Johnson R. and Koehler, K. (1992) “The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure” *Language Learning* Vol. 42/4: 529–555.
- Bresnahan, M. J., Ohashi, R., Nebashi R., Liu W. Y. & Shearman. S.M. (2002) “Attitudinal and affective response toward accented English”. *Language and Communication* 22: 171–185
- [Common European Framework of Reference for Languages: Learning, Teaching, Assessment](#), (2001) Cambridge: CUP
- Cutler, A., A. Weber, R. Smits & N. Cooper (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* 116: 3668–3678

- Delamare, T. (1996) The importance of interlanguage errors with respect to stereotyping by native speakers in their judgements of second language learners' performance. *System* 24: 279–297
- Fayer, J. M. & Krasinski, E (1987) Native and nonnative judgements of intelligibility and irritation. *Language Learning* 37: 313–325
- Galloway, V. B. (1980) Perceptions of the communicative effects of errors in Spanish. *Modern Language Journal* 64: 428–453
- Hieke A.E. (1984) "Toward listener strategies for decoding fluent speech", *IRAL - International Review of Applied Linguistics in Language Teaching*. Volume 28, Issue 3: 221–256
- Hughes, A. & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal* 36: 175–182
- Jenkins J. (2000) *The Phonology of English as an International Language*, Oxford: OUP
- Johansson, S. (1978). "Studies in error gravity: native reactions to errors produced by Swedish learners of English". Gothenburg: *Acta Universitatis Gothoburgensis*
- Koster, C. J. & Koet, T (1993) "The evaluation of accent in the English of Dutchmen". *Language Learning* 43: 69–92
- Leather, J. (1999) "Second language speech research: an introduction". In Leather, J. (ed.) *Phonological Issues in Language Learning*. Oxford: Blackwell: 1–58.
- Mey J. (1998) *Concise Encyclopedia of Pragmatics*, Oxford: Elsevier Science.
- Munro, M. J. & Derwing, M. T (1995). "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners". *Language Learning* 45: 73–97
- Piazza, L.G. (1980) "French tolerance for grammatical errors made by Americans". *Modern Language Journal* 64: 422–427
- Rogerson-Revell, P. "Pronunciation matters: using English for international business communication", paper presented at the Amsterdam Conference on Pronunciation, 2010
- Ryan, E. B. (1983). "Social psychological mechanisms underlying native speaker evaluations of non-native speech". *Studies in Second Language Acquisition* 5: 148–159
- Ryan, E. B., Carranza M.A. & Moffie, R. (1975) "Mexican American reactions to accented English". In J. W. Berry & W. J. Loaner (eds.) *Applied Cross-Cultural Psychology*. Amsterdam, Swets & Zeitlinger: 174–178
- Ryan, E. B. & Sebastian, R.J. (1980) "The effects of speech style and social class background on social judgements of speakers". *British Journal of Social and Clinical Psychology* 19: 229–233.
- Schairer, K. E. (1992) "Native speaker reaction to non-native speech". *Modern Language Journal* 76: 309–319
- Sheorey, R. (1986) "Error perceptions of native-speaking and non-native speaking teachers of ESL". *ELT Journal* 40, 306–312
- Trask, R. (1996) *A Dictionary of Phonetics and Phonology*. London: Routledge
- Upshur J.A. & Turner C.E. (1995) "Constructing rating scales for second language tests", [ELT Journal](#) Volume 49/1: 3-12
- van den Doel R. (2006) *How Friendly are the Natives? An Evaluation of Native-speaker Judgements of Foreign-accented British and American English*, Utrecht: LOT

david.horner@ensae-paristech.fr

David Horner has worked in oral testing since the 1980s. Currently responsible for Cambridge ESOL oral examiners in France and Luxemburg and head of the Language Department at ENSAE ParisTech, David's book on oral testing, "Le CECRL et l'évaluation de l'oral", published by Belin in France, has just appeared.

Linking Certification to the CEFR: Do we need standard-setting?

Brian North, EAQUALS / Eurocentres Foundation, Switzerland



Introduction

In this paper I will first recapitulate the purpose and relevance of the Common European Framework of Reference (CEFR: Council of Europe 2001) and the procedures recommended for relating tests to the CEFR in the Manual published in 2009 after substantial piloting since the 2003 preliminary edition (Council of Europe 2003; 2009). After that I will briefly mention the scheme inspired by the Manual that we recently introduced in EAQUALS (European Association for Quality Language Services www.eaquals.org) for member schools that wish to issue EAQUALS Certificates of Achievement to their learners. Then I will discuss certain aspects related to criterion-referenced assessment (CR) and standard-setting and how this concerns relating assessments to the CEFR.

Conventional standard-setting, with a panel estimating item difficulty level in order to set the cut-score for pass/fail or different grades in a test, seems to be considered essential in EALTA. Eli Moe, for example, in her paper at the EALTA standard setting seminar starts off:

“Although everyone agrees that standard-setting is a must when linking language tests to the Common European Framework of Reference (CEFR), we hear complaints about the fact that standard-setting is expensive both in respect to time and money. In addition it is a challenge to judges not only because the CEFR gives little guidance on what characterises items mirroring specific levels, but also because time seldom seems to increase individual judges’ chances of success in assigning items to CEFR levels.” (Moe 2009: 131)

Neither I nor Neil Jones nor John De Jong, to name but three people present, would agree that panel-based standard-setting is a “must” when linking tests to the CEFR. The preliminary, pilot version of the Manual presented only one method of panel-based standard-setting, the so-called DIALANG or “basket” method (explained later). The preliminary Manual also made it clear that it was perfectly feasible to jump from the specification phase direct to empirical, external validation

without bothering with panel-based standard-setting at all. It recommended using the judgements of CEFR-trained teachers to do so and presented box plots and bivariate decision tables provided by Norman Verhelst, the Cito statistical expert from the DIALANG project, as useful tools in that process. When I met Norman at the first meeting of the Manual group, we in fact had a one hour discussion in which I expressed my difficulty in buying the idea that someone with his experience of Item Response Theory (Rasch modelling, henceforth IRT) could seriously believe that such guesstimation by panels could work. After all there had been doubts since the 1970s and there are effective alternatives, to which we will return later.

The Purpose of the CEFR

But first let us remind ourselves what the CEFR is all about. Published in 2001 after a period of piloting, it consists of a descriptive scheme, common reference points expressed as six proficiency levels, descriptor scales for many aspects of that descriptive scheme, advice on curriculum scenarios and considerations for reflection. The aim of the CEFR is to stimulate reflection on current practice and to provide the common reference levels to facilitate communication, comparison of courses and qualifications, and personal mobility. The way it is expressed (Council of Europe 2001: 178) is that the CEFR can be of help:

for the specification of the content of tests and examinations:	<i>What is assessed</i>
for stating the criteria to determine the attainment of a learning objective:	<i>How performance is interpreted</i>
for describing the levels of proficiency in existing tests and examinations, thus enabling comparisons to be made across different systems of qualifications:	<i>How comparisons can be made</i>

Before the CEFR, there was a practical “Tower of Babel” problem in making sense of course certificates and test scores. A teacher, school or examination body would carry out a test and report a result in their own way as “19,” “4.5,” “516,” “B,” “Good,” etc. It is no exaggeration to say that 20 years ago a teacher of Spanish in a secondary school in southern France, a teacher of French to Polish adults and a teacher of English to German businessmen would have taken 10-20 minutes to establish any common ground for a real discussion. The CEFR labels help. But the old jungle also masked a theoretical problem: of relating assessment results to real world practical language ability. As Jones et al (2010: 230) point out, the CEFR and Manual help

language testers to address this central concern of criterion-referenced assessment. The CEFR promotes an “action-oriented approach:” seeing the learner as a language user with specific needs, who needs to *act* in the language. The CEFR descriptor scales can provide the vertical continuum of real life ability needed as an external criterion for valid criterion-referenced assessment. This point is returned to later in the paper.

However, it is important to remember that the prime function of the CEFR is to encourage reflection on current practice. It offers a heuristic model, not a panacea. This fact was recognised in various articles in the special issue of the journal *Language Testing* on the CEFR:

... in Chapter 2, the principal intended uses of the CEFR are made clear: though arbitrary, proficiency descriptions and scales provide an essential heuristic for understanding and communicating about language learning and use, and such a heuristic is needed in a contemporary Europe that seeks to promote mutual understanding, tolerance and knowledge of its rich linguistic and cultural diversity.” (Norris 2005: 400)

“It is essential that the CEFR is not seen as a prescriptive device but rather a heuristic, which can be refined and developed by language testers to better suit their needs”. (Weir 2005: 298)

Secondly it is important to privilege profiling over levelling. The label A2 is always a convenient summary of a complex profile. Demonstrating that two tests are A2 does not entail a claim that the two tests are equivalent or interchangeable. The philosophy of the CEFR is to use the descriptor scales to *profile* the assessment under study, as in the example from the Manual (Council of Europe 2003: 63; 2009: 33) given in Table 1. This illustrates a Belgian examination for immigrants in Dutch as a foreign language. The vertical axis on the left represents the CEFR levels. The horizontal axis shows overall language proficiency plus the CEFR categories covered – both in terms of communicative language activities and in terms of aspects of language competence. The categories chosen to illustrate the coverage of different exams can and should be different.

C2								
C1								
B2.2								
B2								
B1.2								
B1								
A2.2								
A2								
A1								
Overall	Listening	Reading	Social Conversation	Information Exchange	Notes, Messages and Forms	Sociolinguistic	Pragmatic	Linguistic

Fig. 1 CEFR Manual: Form A23: Graphic Profile of the Relationship of the Examination to CEFR Levels (Example)

Profiling means that exams do not need to be compared directly to each other, or claim to be exact equivalences of each other, they can be related to each other through their CEFR profile.

Procedures for relating assessments to the CEFR

The CEFR levels are intended for common reference. It is clear that what exactly is meant in practice by a set of verbally defined levels of proficiency like the CEFR Common Reference Levels cannot be entirely separated from current process of implementation, training workshops, calibration of illustrative samples, adaptation of CEFR descriptors, and linking of tests to the CEFR. However, the levels are not intended for a free-for-all under which people define their own interpretation of them. As was emphasised at the intergovernmental Language Policy Forum held in 2007 to take stock of implementation of the CEFR, the levels should be applied responsibly, especially in any alignment of national systems and international certificates to them. The Manual recommends four sets of procedures for linking to the CEFR: familiarisation,

specification, standardisation and validation. In the context of operational school assessment, one could also put a special emphasis on moderation techniques to limit and/or adjust for subjectivity in assessments by teachers.

Familiarisation with the CEFR levels through training and awareness-raising exercises is always necessary as people tend to think they know the levels without consulting the descriptors or official illustrative samples. Instead they often associate the CEFR levels with levels they are already familiar with. Familiarisation exercises normally involve descriptor sorting tasks, but the most useful, initial form of familiarisation is to see the levels in action – in video sequences such as those available online for English, French, Spanish, German and Italian on http://www.ciep.fr/en/publi_evalcert/dvd-productions-orales-cecrl/index.php.

Specification entails defining the coverage of the course or examination in relation to the CEFR descriptor scales, both in terms of the curriculum and in terms of the assessment tasks and criteria used to judge success in them. This involves selecting the communicative activities, perhaps guided by descriptor scales in CEFR Chapter 4 (summarised in CEFR Table 2), designing tasks and writing items. Valid assessment requires the sampling of *a range of relevant discourse*. For speaking, this means combining interaction (spontaneous short turns) with production (prepared long turns). For writing it may mean eliciting written-spoken language (interaction: email, SMS, personal letter) as well as prose (production: essay, report). For listening and reading it may mean some short pieces for identifying specific information and one or two longer pieces for detailed comprehension.

The formulation of criteria may or may not be related to descriptors in CEFR chapter 5 (summarised in CEFR Table 3), but criteria should be balanced in terms of extent of knowledge and degree of control, and of linguistic competence and pragmatic competence. The assessment instrument might be a single grid of categories and levels like CEFR Table 3, especially for standardisation training or a programme in which teachers teach classes at different levels. On the other hand it might focus only on the target level, with one descriptor per chosen category, as shown with a simple example in Table 2. The advantage of this approach is the ease with which the criteria can be explained to learners. This makes it easier to highlight the qualities and competences they must acquire for communicative success, rather than just focusing on lists of things they “can do” or just grammar and vocabulary.

		Test (Item bank)					Total
		A1 (1)	A2 (2 & 3)	B1 (4 & 5)	B2 (6 & 7)	C1 (8 & 9)	
Criterion (Teachers)	A1 (1)	4	1				5
	A2 (2 & 3)		14	4			18
	B1 (4 & 5)		5	13	2		20
	B2 (6 & 7)			3	16		19
	C1 (8 & 9)				3	3	6
	Total	4	20	20	21	3	68

Fig. 2 A CEFR Manual “Decision Table” for validation of cut scores on a Eurocentres item bank for German (North 2000b)

Standardisation involves training in a standard interpretation of the levels, using the illustrative samples provided for that purpose, and secondly the transfer of that standardised interpretation to the benchmarking of local reference samples. It is important that one does not confuse these two things. In standardisation, participants are trainees being introduced to or reminded of the levels, the criteria, the administration procedures etc. There is external authority represented by the workshop leader, the official criteria and the calibrated samples. Standardisation training is not an exercise in democracy. The right answer, in terms of standardising to an interpretation of the levels held in common internationally, is not necessarily an arithmetic average of the opinions of those present, especially if they all come from the same school or pedagogic culture. This is a tricky issue which needs to be handled delicately. Personally I have found it simplest to start by showing a video, allowing group discussion, handing out the documentation and then animating a discussion of why (not whether) the learner is A2. The next stage can have group discussion reporting views to plenary, and finally individual rating – checked with neighbours.

In benchmarking, on the other hand, participants are valued, trained experts (although quite possibly the same people who did the standardisation training in the morning!) Here it is important to record individual judgements before they are swayed by over-dominant members of

the group in discussion. Ideally the weighted average of the individual judgements, preferably corrected for inconsistency and severity/lenience with the IRT program FACETS, (Linacre 1989; 2008) would yield the same result as the consensus reached in discussion. This was the preferred method in the series of benchmarking seminars that produced illustrative video samples (North and Lepage 2005; Jones 2005).

Benchmarking used for estimating the difficulty of *test items* (rather than the ability of learners shown in video clips) is a form of standard-setting. As I commented at the beginning of this paper, it often seems to be referred to in EALTA as if it were the *only* form of standard-setting. Yet in fact, as Ellie Moe commented, the process is very prone to error and the only real advantage of such methods is that some can be employed by teachers without much need for statistical knowledge. Examination institutes should be relating their *reporting scale* to the CEFR so as to guarantee the link over different test administrations; they should not be relating particular items or one particular test form on the basis of the views of one particular panel. People can choose whether or not to use a benchmarked video clip – they can prefer one illustrative sample to another. But when this technique is used as *the* way to relate a test to a proficiency scale it suggests a considerable act of faith in the ability of the panel to perform a very indirect and difficult judgement. This is a point returned to later in the paper.

Moderation involves counteracting subjectivity in the process of rating the productive skills. Even after standardisation training has been implemented, moderation will always be necessary. Some assessors can be quite resistant to training and the effects of the standardisation also start to wear off immediately after the training anyway. In addition, some assessors persist in using personal concepts rather than the official criteria as their reference, many are unconsciously over-influenced by one criterion (e.g. accuracy or pronunciation), and most refuse to give a top or bottom grade (= error of central tendency). Moderation techniques can be divided into collective and quality control techniques. Collective techniques involve some form of double marking, perhaps of a structured sample of candidates (e.g. every 5th consecutive candidate, or (after rank ordering) the top three, middle three, and bottom three candidates). Rather than live double marking, recordings might be sent to “chief examiners” for external monitoring. Administrative quality control techniques may involve studying collateral information on the candidates on the one hand, or developing progress norms from representative performance samples sent to the “chief examiners” on the other. Such norms can then be used to identify classes whose grades differ significantly from the norm, for further investigation.

These grades might be due to an unusually good/bad teacher or an unusually strong/weak class – but it is worth following up.

EAQUALS Scheme

These techniques (familiarisation, specification, standardisation, moderation) have recently been operationalised in a scheme for EAQUALS-accredited language education providers to issue EAQUALS CEFR Certificates of Achievement to learners at the end of a course. The scheme requires the school to send the following materials for inspection by an expert panel and the school's assessment system is then checked in practice during the 3-yearly EAQUALS external inspections:

- curriculum and syllabus documents with learning objectives derived from the CEFR
- a coherent description of the assessment system
- written guidelines for teachers
- CEFR-based continuous assessment instruments
- sample assessment tasks, tests, guidelines
- CEFR-based criteria grids
- a set of locally recorded, CEFR-rated samples to be double checked by an EAQUALS expert panel
- samples of individual progress records
- content and schedule of staff training
- details of moderation techniques employed

Validation involves two aspects: *internal validation* of the intrinsic quality of the assessment and *external validation* of the claimed link to the vertical continuum of real-life language ability operationalised in the CEFR descriptor scales. For reasons of space I shall only discuss the latter, since the entire language testing literature concerns the former. Many of the moderation techniques referred to above are simple forms of external validation: the fundamental principle is to exploit collateral information and independent sources of evidence. The advice in the Council of Europe's Manual to use two independent methods of setting the "cut scores" between levels, and then if necessary use a cyclical process of adjusting the "cut scores" and examining them with a "decision table" like that shown in Table 3 in order to arbitrate between the two provisional results. The table shows a low-stakes worked example cited in the Manual (Council of Europe 2009: 111–3); here the pattern was very regular with 73.5% matching classifications, so no

correction from the provisional cut scores set for the item bank on the basis of item writer intention seemed necessary.

	Candidate A				
RANGE & PRECISION: Can talk about familiar everyday situations and topics, with searching for the words; sometimes has to simplify.	1	2	3	4	5
ACCURACY: Can use some simple structures correctly in common everyday situations.	1	2	3	4	5
FLUENCY: Can participate in a longer conversation about familiar topics, but often needs to stop and think or start again in a different way	1	2	3	4	5

Table 3: Assessment at One Level

This contrastive technique can be exploited in many different ways, for example: contrasting the original claim based on item writer intention versus a result from formal standard setting; contrasting the results from two independent standard-setting panels; contrasting the results from two different standard-setting methods (e.g. between a test-centred and a candidate-centred method), and finally confirming the result from standard-setting (or the original claim based on item writer intention) with a formal external validation study. The external criterion could be operationalised in CEFR-related examination results for a school assessment or in the judgements of CEFR-trained class teachers for a test under study. In fact many of the Manual case studies recently published (Martyniuk 2010) did successfully use two methods in order to confirm their claim to CEFR linkage. Both the ECL study (Szabo 2010) and TestDaf study (Kecker & Eckes 2010) contrasted original item-writer intention with formal standard-setting; both the City & Guilds study (O’Sullivan 2010) and the ECL study (Szabo 2010) contrasted the mean average difficulty of their own items with that of the illustrative items; both TestDaf study (Kecker & Eckes 2010) and the Bilkent COPE study (Thomas & Kantarcioğlu 2009; Kantarcioğlu et al 2010) contrasted panel-based standard-setting results with external teacher judgements of the candidates in relation to CEFR descriptors. Finally the Surveylang study (Verhelst 2009) contrasted results from a sophisticated data-based, panel “bookmark method” (Council of Europe 2009: 82–3) with such external teacher CEFR judgements. Both the Pearson Test of English – Academic (De Jong 2010) and the Oxford On-line Test (Pollitt 2009) contrasted item writer intentions with external teacher judgements.

In contrast to these sensible approaches, Cizek and Bunch (2007), the current US text book on standard-setting, explicitly advise against using two methods of standard-setting, because these might yield different results. They state that “a man with two watches is never sure” and “use of

multiple methods is ill advised” (Cizek and Bunch 2007: 319-20). Yet replication is the basis of western academic thought: if you cannot replicate a result you do not have a result. Good practice would dictate corroboration of what is, for a high stakes test, an important decision that will affect many people’s lives.

Criterion-referencing and Standard-setting

As discussed above, validating the relationship of a test to the CEFR requires what is technically known as “linking” to the continuum of ability acting as the criterion. Criterion-referenced assessment (CR) places persons on that continuum independent of the ability of others. In our current discussion, the validated CEFR descriptor scale provides that continuum. CR was developed by Robert Glaser in a seminal article from which the crucial passage is the following:

“Along ... a continuum of attainment, a student’s score on a CR measure provides explicit information as to **what the individual can and cannot do.**

CR measures indicate (...) the correspondence between what an individual does and **the underlying continuum of achievement.** Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is **independent of reference to the performance of others.**” (Glaser 1963: 519–20)

This is not at all where the conventional, US-style standard-setting represented by Cizek & Bunch (2007) is coming from. Because CR started in the 1960s in the US at almost exactly the same time as the behaviourist instructional objectives movement, the two concepts unfortunately merged in setting the “performance standard” for “mastery” in the US “minimum competence” approach (Glaser R. 1994a: 6; 1994b: 9; Hambleton 1994: 22). Over time, that performance standard – which is a norm: a definition of what it might be reasonable to expect from a newly qualified professional, or from a 3rd year high school student in a specific subject in a specific context – became confused with the criterion – which is supposed to be the continuum of real-world ability. “Standard-setting” then became the process of setting the pass/fail norm for minimum competence in a multiple-choice test for a given body of knowledge in the subject concerned. Since it was the experts (panel of expert nurses; committee of 3rd year teachers) that defined that body of knowledge, they were also in a position to give an authoritative judgement on whether the test was “fair.” The “fairness” relates to what people feel it is reasonable to expect from a specific cohort of candidates in relation to the closed domain of knowledge concerned. Whether an individual’s result is considered to be good or bad therefore depends

entirely on how that result relates to the score set as the expected norm for the group of whom they are part. This is fair enough, but it is neither criterion-referenced assessment nor proficiency assessment, in the sense in which the word is used in the expression “language proficiency.” There is no relationship to an external criterion: the continuum of language ability.

As Jones (2009:36) pointed out in his paper at the EALTA standard-setting seminar there really is almost nothing in common between setting such a pass norm for a closed domain of knowledge on the one hand and linking a language test to the continuum of language proficiency articulated by the CEFR on the other. In addition, as Reckase (2009: 18) suggested in his EALTA presentation, panel-based methods were not designed for the multiple cut scores necessary in linking results to different language proficiency levels; there is an inevitable dependency between the decisions.

Twenty-three out of the out of the twenty-six articles on case studies of relating tests to the CEFR in Martyniuk (2010) and Figueras and Noijons (2009) took a panel-based standard-setting approach, mostly citing Cizek and Bunch (2007) – though as mentioned earlier several did replicate their findings with a second method. This demonstrates the extent to which many language testers and many people involved in linking assessments to the CEFR are not aware of the confusion between criterion-referencing and mastery learning described above, nor that panel-based standard-setting is a norm-referencing technique, nor that it is not innately suitable as a means to set multiple cut-scores on a test. Nor are many language testers aware that there is 30 years of literature attesting to the fact that such panel-based standard-setting is flawed even within its own context (e.g. Glass 1978: 240 – 42; Impara and Plake 1998: 79).

This fact has recently been rediscovered (Kaftandjieva 2009) in an EALTA context in the evaluation of the so-called “basket method” used in the DIALANG project and recommended in the preliminary, pilot version of the Manual. The basket method is one of the many modified Angoff methods and asks the panel member to decide which “basket” (A1, A2 etc) to put the item in, by posing and answering a question like “*At which CEFR level will a candidate first be able to answer this question correctly?*” Other panel-based methods feed data to panellists between rounds, usually on item difficulty (facility values or IRT theta values) and then on “impact” (how many people would fail if we said this) and as Kaftandjieva (2009: 30) indicates such a modified basket method works much better. But all these endless Angoff variants appear to me to be really just exercises in damage limitation.

Calibrating to a common criterion

Best practice in linking a high stakes test to the CEFR involves *calibrating* the scale behind the test or suite of tests to the external criterion provided by the CEFR descriptor scale. This is technically called vertical scaling or vertical equating and is the main advantage offered by IRT (Item Response Theory). Simple introductions to IRT are offered by Baker (1997), McNamara (1996), and Henning (1991). Cizek & Bunch (2007) devote just 7% of their text to the issue of standards at different stages on a continuum of ability – only to then reject the concept. They discuss what they describe as “vertically moderated standard-setting” (VMSS) which is a way of smoothing out infelicities when stringing together a series of norms for different school years, each determined independently by standard-setting panels. They conclude that “none (*of the VMSS methods*) have any scientific or procedural grounding to provide strong support for its use” (Cizek & Bunch 2007: 297). Vertical scaling to a continuum of ability (=IRT) they reject out of hand on the basis of a study by Lissitz and Huynh (2003). Yet Lissitz and Huynh cite 6 specific reasons why vertical scaling with IRT was inappropriate *for their context*. These are briefly summarized and commented on below. None of them apply to the context of relating language assessments to the CEFR.

1. It is only suitable for subjects like reading or math.

Second language proficiency is a subject like reading or math; it too has a clear continuum of ability.

2. A common dimension across school years doesn't capture year-specific instructional expectations

It can do if care is taken to include items in a live item bank only when their “level” is the same from the point of view of both curriculum considerations and empirical difficulty. That is the principle of a good item bank, and one that was proposed by the Dutch Construct Group.

3. It confounds content changes with method changes over school years

This would be a problem if one were foolish enough to use a single item bank for all educational sectors or for all years of a sector, when substantial cognitive development over the years concerned changed the nature of the learning process and hence of the tasks set. One would also not expect all skills to “fit” into the same bank anyway. The CEFR linking process offers a way of relating several different item bank scales to the common metric, as already done, for example, by Cambridge ESOL.

4. The assumption of equal intervals on the scale means that comparison of growth

(between different groups; between the same group at two points in time) cannot satisfactorily be made.

The CEFR descriptor scale had almost equal intervals in the original research (North 2000a: 273–4) if the level “Tourist” below A1 is included as a pre-level for A1, and if C1 and C2 are seen as one broad band. The empirical scale had 10 almost equidistant bands: Tourist, A1, A2, A2+, B1, B1+, B2, B2+, C1, and C2. It is difficult to understand why an assumption of equal distance (not usually made in discussion of CEFR levels anyhow) would make comparisons more difficult. On the contrary, I would have thought that a common metric would make it easier; that is the whole point of a common metric.

5. Performance and learning are multidimensional. The single dimension required by a scale encourages simplifications and the loss of the very insights assessments were carried out to illuminate.

This is an argument against restricting assessment to what can be tested with test items/descriptors that can be calibrated into an item bank. However, the CEFR descriptors provide a rich description of features and competences that occur at different levels that can inform criteria used for feedback on performance and learning. Calibrated tests and descriptors can also be supplemented by descriptors giving feedback on aspects of sociocultural, intercultural and strategic competences that probably cannot be scaled at all.

6. Vertical scaling with IRT is a technically difficult task. Adjustments need to be made to smooth the results (Camilli 1999 – and also 1988)

This refers to the need for a very good IRT anchoring design and for quality control in relation to scale distortion. As Camilli discovered, IRT scales distort in the top 20% and the bottom 20% of the scale. With overlapping scales from different tests anchored together through common items in a conventional IRT “missing data design,” this can lead to an exaggerated overlap between the different test forms on the common scale. The problem can be corrected by eliminating items or people scoring above 80% or below 20% (Jones 1993; North 2000a); it can be reduced by using the OPLM IRT model developed by Cito (accurate between 90% and 10%), and it can be avoided entirely by anchoring all test forms 50% upwards and 50% downwards to adjacent tests, thus cancelling out the distortion (De Jong, personal communication).

Linking Assessments

There is actually a literature on linking assessments and I find it surprising that only one of the twenty-six articles in Martyniuk (2010) and Figueras & Noijons (2009) referred to it (Maris 2009), though another did report equating the test logit scale and logit scale of the CEFR descriptors

(David 2010). Angoff's (1971) article was part of this literature, entitled "Scales, Norms and Equivalent Scores." The so-called Angoff standard-setting method was in fact a remark in a footnote. Before becoming involved in the development of the Manual I wrote a modest article entitled "Linking language assessments: an example in a low stakes context" (North 2000b). It described the way over the years various people in Eurocentres, a chain of schools teaching languages where they are spoken, had addressed the question of equating tests and linking them to the Eurocentres scale of proficiency, a pre-cursor of the CEFR descriptor scale. I had read most of the then standard-setting literature in bibliographic research before the development of the CEFR descriptor scale, but didn't see how panel-based, judgemental methods were relevant to a common framework scale of levels, except for rating spoken and written samples. Even there it seemed clear that many-faceted IRT scaling was needed to handle inconsistency and subjectivity (Linacre 1989; 2008) as applied in the CEFR research project (North 2000a) and in calibrating the CEFR illustrative spoken samples (North & Lepage 2005, Breton et al 2007).

I certainly think that the experience of participating in standard-setting seminars is a very enriching one. It is very valuable awareness-raising and training for a team of test developers and item writers to consciously evaluate and judge the difficulty of items, and then be confronted with empirical data on item difficulty. As Moe (2009: 137) suggests this process may also help make the levels more concrete by teasing out their criterial features. But why use such a qualitative, guesstimation approach to set cut-scores? There is a data-based alternative that exploits vertical scaling and the judgements of CEFR-trained teachers. As mentioned above, the technique has been used in several CEFR linking projects (Oxford Online Placement Test: Pollitt 2009; Pearson Tests of English: De Jong 2010; the UK Languages Ladder project: Jones et al 2010, and the European Survey of Language Competence: Verhelst 2010). The technique is explained in the "Further Material" (North & Jones 2009) provided to accompany the Manual. This is buried in the small print on the Council of Europe's website (www.coe.int/lang), sandwiched between the link to the Manual text and the link to the Reference Supplement. I thoroughly recommend it to you.

Conclusion

The CEFR is a useful heuristic tool, but it is not the answer to all problems. It is an inspiration not a panacea. It needs further exemplification, as in the banks of illustrative descriptors and samples on the Council of Europe's website. It requires the elaboration of content for different

languages, as in the “Reference Levels” for German, French, Spanish and Italian and in the recently published British Council/EQUALS *Core Inventory for General English* (North et al 2010), not to mention the much awaited Cambridge-led corpus-based project *English Profile*.

There is no “official” way of linking tests to the framework. There is a Manual; there is what is in effect a minority report from the Manual team (“Further Material”), and there is a further body of work undertaken by ALTE and by Cambridge ESOL (e.g. see Khalifa and Weir 2009, Khalifa et al 2010).

Fundamentally the CEFR, the Manual, the Further Material, the Reference Levels, the descriptor banks and the illustrative samples are all reference tools to be critically consulted, not things to be “applied.” The boxes at the end of each CEFR chapter invite users to reflect on their current practice and the way in which it relates to what is presented in the CEFR. The authors of many of the case studies published in Martyniuk (2010) on relating tests to the CEFR state that the process of undertaking the project led them into such a process of reflection and reform. It is such a process that the CEFR was designed to stimulate.

References

- Angoff, W.H. (1971): Scales, Norms and Equivalent Scores, in Thorndike, R.L. (ed.) *Educational Measurement*, Washington D.C. American Council on Education, 508–600.
- Baker, R. (1997): *Classical Test Theory and Item Response Theory in Test Analysis. Extracts from: An Investigation of the Rasch Model in Its Application to Foreign Language Proficiency Testing*. Language Testing Update Special Report No 2.
- Breton, G., Lepage, S. and North, B. (2008): *Cross-language Benchmarking Seminar to Calibrate Examples of Spoken Production in English, French, German, Italian and Spanish with Regard to the Six Levels of the Common European Framework of Reference for Languages (CEFR)*. CIEP, Sèvres, 23-25 June 2008. Report. Strasbourg: CIEP/Council of Europe, www.coe.int/lang
- Camilli, G. (1988): Scale Shrinkage and the Estimation of Latent Distribution Parameters. In: *Journal of Educational Statistics* 13, 3, 227–241.
- Camilli, G. (1999): Measurement Error, Multidimensionality, and Scale Shrinkage: A Reply to Yen and Burket. *Journal of Educational Measurement*, 36, 73-78.
- Cizek, G.J. and Bunch, M.B. (2007): *Standard Setting: a Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks: Sage.
- 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)* DGIV/EDU/LANG (2003) 5, Strasbourg, Council of Europe.
- Council of Europe (2009): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*, Strasbourg: Council of Europe www.coe.int/lang.

- Dávid, G. (2010): Linking the General English Suite of the Euro Examinations to the CEFR: A Case Study Report, in Martyniuk, W. (ed.), 177–203.
- De Jong (2010): *Aligning PTE Academic Score Results to the Common European Framework of Reference for Languages*. Accessible at <http://pearsonpte.com/research/Documents/AligningPTEtoCEF.pdf>
- Figueras, N. and Noijons, J. (eds.) (2009): *Linking to the CEFR Levels: Research Perspectives*. Arnhem, Cito-EALTA.
- Glaser, Robert (1963): Instructional Technology and the Measurement of Learning Outcomes: Some questions, in *American Psychologist* 18, 8, 519–521.
- Glaser, R. (1994a): Instructional Technology and the Measurement of Learning Outcomes: Some Questions. *Educational Measurement: Issues and Practice*, 13, 4, 6–8.
- Glaser, R. (1994b): Criterion-referenced Tests: Part 1. Origins, *Educational Measurement: Issues and Practice*, 13, 4, 9–11.
- Glass, Gene V. (1978): Standards and Criteria. In: *Journal of Educational Measurement* 15, 4, 237–261.
- Hambleton, R.K. (1994): The Rise and Fall of Criterion-referenced Measurement? *Educational Measurement: Issues and Practice*, 13, 4, 21–26.
- Henning G. (1987): *A Guide to Language Testing*, Newbury House.
- Impara, J.C. and Plake, B.S. (1998): Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard-setting Method. *Journal of Educational Measurement*, 35, 1, 69–81.
- Jones, N. (2009): A Comparative Approach to Constructing a Multilingual Proficiency Framework Constraining the Role of Standard-setting, in Figueras, N. and Noijons, J (eds.), 35–44.
- Jones, N. (1993): *An Item Bank for Testing English Language Proficiency: Using the Rasch Model to Construct an Objective Measure*, PhD thesis, University of Edinburgh.
- Jones, N. (2005): *Seminar to Calibrate Examples of Spoken Performance, CIEP Sèvres, 02-04.12.2004. Report on analysis of rating data, final version*. March 1, 2005. www.coe.int/lang.
- Jones, N., Ashton, K. and Walker, T. (2010): Asset Languages: A Case Study of Piloting the CEFR Manual, in Martyniuk, W. (ed.), 227–248.
- Kaftandjeva, F. (2009): Basket Procedure: The Breadbasket or the Basket Case of Standard-setting Methods? in Figueras, N. and Noijons, J. (eds.), 21–34.
- Kantarcioglu E, Thomas C, O'Dwyer J. and O'Sullivan, B. (2010): Benchmarking a High-Stakes Proficiency Exam: The COPE Linking Project, in Martyniuk, W. (ed.), 102–118.
- Kecker, G. and Eckes, T. (2010): Putting the Manual to the Test: The TestDaf–CEFR Linking Project, in Martyniuk, W. (ed.), 50–79.
- Khalifa, H. and Weir, C. (2009): *Examining Reading: Research and Practice in Assessing Second Language Reading*, Cambridge, Cambridge University Press, Studies in Language Testing 29.
- Khalifa, H, French A. and Salamoura, A. (2010): Maintaining Alignment to the CEFR: The FCE Case Study, in Martyniuk, W. (ed.), 80–102.
- Linacre, J.M. (1989): *Multi-faceted Measurement*, Chicago, MESA Press.
- Linacre, J.M. (2008): *A User's Guide to FACETS, Rasch Model Computer Program*, ISBN 0-941938-03-4. www.winsteps.com.
- Lissitz, R.W. and Huynh, H. (2003): Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability. *Practical Assessment, Research & Evaluation*, 8, 10, <http://pareonline.net/getvn.asp?v=8&n=10>.
- Maris, G. (2009): Standard-setting from a Psychometric Point of View, in Figueras, N. and Noijons, J. (eds.), 59–66.
- Martyniuk, W. (ed.) (2010): *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual*, Cambridge, Cambridge University Press, Studies in Language Testing 33.

- MacNamara, T. (1996): *Measuring Second Language Performance*, London and New York: Longman.
- Moe, E. (2009): Jack of More Trades? Could Standard-setting Serve Several Functions? in Figueras, N. and Noijons, J. (eds.), 131–8.
- Norris, J.M. (2005): Common European Framework of Reference for Languages: Learning, Teaching, Assessment, *Language Testing* 22, 3, 399–406.
- North, B. (2000a): *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang.
- North, B. (2000b): Linking Language Assessments: An Example in a Low Stakes Context, *System* 28, 555–577.
- North, B. and Jones, N. (2009): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*, Strasbourg: Council of Europe, www.coe.int/lang
- North, B. and Lepage, S. (2005): *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, Report. Strasbourg: Council of Europe, www.coe.int/lang
- North, B., Ortega Calvo, A. and Sheehan, S. (2010): *British Council –EAQUALS Core Inventory for General English*, London, British Council/EAQUALS, ISBN 978-086355-653-1. Downloadable from www.teachingenglish.org.uk and www.eaquals.org
- O’Sullivan, B. (2010): The City & Guilds Communicator Examination Linking Project: A Brief Overview with Reflections on the Process, in Martyniuk, W. (ed.) 33–49.
- Pollitt, A. (2009): *The Oxford Online Placement Test: The Meaning of OOPT Scores*. Oxford: Oxford University Press, downloaded from www.oxfordenglishtesting.com.
- Reckase (2009): Standard-setting Theory and Practice: Issues and difficulties, in Figueras, N. and Noijons, J. (eds.), 13–20.
- Szabo, G. (2010): Relating Language Examinations to the CEFR: ECL as a Case Study, in Martyniuk, W. (ed.), 133–44.
- Thomas & Kantarcioğlu (2009): Bilkent University School of English Language COPE CEFR Linking Project, in Figueras, N. and Noijons, J. (eds.), 119–124.
- Verhelst, N. (2009): Linking Multilingual Survey Results to the Common European Framework of Reference, in Figueras, N. and Noijons, J. (eds.), 45–58.
- Weir, C. (2005): Limitations of the Common European Framework for Developing Comparable Language Examinations and Tests’ *Language Testing*, 22, 3, 281–300.

Brian North was co-author of the CEFR, developer of the CEFR descriptor scales and coordinator of the CEFR Manual group. He is Head of Academic Development at Eurocentres, the language school foundation that has been an NGO to the Council of Europe since 1968. He is Vice-Chair of EAQUALS and co-ordinates the EAQUALS Curriculum and Assessment SIP.

Designing fair writing testing tasks

Vasso Oikonomidou, MA, PhD Candidate, University of Athens, Greece



Introduction

The issue of validity and fairness in task design for assessment purposes has been considered crucial in the field of language testing (Bachman & Palmer, 1996; CEFR, 2001). Especially when it comes to considering the case of high stakes exams, the demand for the design of valid and ‘fair’ tasks, becomes even greater. It is, therefore, important that item writers and test designers consider the different variables which affect test validity and fairness and bear in mind that different types of tasks may create different kinds of difficulty to test-takers coming from a different social or educational background (Purves, 1992).

This paper aims at offering some insights towards the development of a construct validity framework for the design of ‘fair’ tasks for testing writing. It comprises both a theoretical and a practical part. The theoretical part sheds some light on the concept of fairness and validity in testing, which also the CEFR is concerned with. The practical part presents two examples of writing tasks from different exam batteries for the illustration of particular points such as target population, task type and task content, which item writers should bear in mind. The issues presented in this paper come from the researcher’s study of the relevant literature as well as findings from on-going research on writing task design in the field of language assessment.

Fairness in language testing: a definition

The fairness of language tests has always been a concern among item writers and test developers. In the related literature the conceptualization of fairness has been viewed from different perspectives (Xi, 2010), which relate to the connection between fairness and validity. Fairness has either been regarded as an independent facet of test quality or as a facet directly linked to validity. Xi (ibid.), for instance, defines fairness as ‘comparable validity for *relevant*

groups that can be identified'. In this sense, a test may be more 'fair' for certain groups of test-takers but not for others. Therefore, to make language tests as fair as possible, it is crucial that item writers and test designers follow certain pre-determined guidelines and illustrative descriptors but also consider the different variables that may affect test-taker performance. In this way, the content validity of the test is also assured.

Content validity: an aspect of validity

The importance of validity in the testing context is discussed in Bachman & Palmer (1996), where 'construct validity' is referred to as one of the six qualities comprising a model of test usefulness. Also, as noted in the CEFR (2001:177), "an assessment procedure can be said to have validity to the degree that it can be demonstrated that what is actually assessed (the construct) is what in the context concerned, should be assessed, and that the information gained is an accurate representation of the proficiency of the candidate concerned". One form of construct validity, which test designers should consider in task design, is content validity. Table 1 shows the different parameters that should be taken into account in relation to testing writing, which this paper focuses on. All the parameters mentioned in Table 1 are equally important for item writers to consider in writing task design. This paper, however, is limited to a discussion of a. target population, b. task type, and c. task content.

Table 1: Content Validity

<p>A. THE TEST</p> <ul style="list-style-type: none"> ➤ PURPOSE ➤ DEFINITION OF CONSTRUCT ➤ TARGET POPULATION ➤ TASK TYPE 	<p>B. TEST SPECIFICATIONS</p> <ul style="list-style-type: none"> ➤ CONTENT ➤ STRUCTURE & TIMING ➤ CRITERIAL LEVELS OF PERFORMANCE
--	---

a. Target population

The target population is one of the variables that should be taken into account by item writers in writing task design as it may affect test performance. In particular, test designers should consider test-takers' age as well as their educational, social and cultural background. For instance, young candidates may not have developed as many skills as adults since they have not been exposed to as many stimuli as adults, which should be taken into account in task

design and the articulation of output expectations. In addition, test-takers who live in urban rather than rural areas have different experiences and different world knowledge. Finally, education plays a significant role in the development of different kinds of literacy.

b. Task type

Task type is reflected in task prompts, which, according to Kroll & Reid (1994), are the stimuli test-takers respond to in testing situations. Kroll & Reid (*ibid.*) refer to three different types of prompts: a. bare prompts, which state the entire task for test-takers, b. framed prompts, which present a situation, and c. text-based prompts, where a text is provided to test-takers as input on the basis of which they are expected to produce their written response, either by summarizing the main points of it or by responding to and interacting with it. The latter task type is also referred to in the literature as 'read-to-write' or 'reading-to-write'.

Each task type places different demands on test-takers. The existence of a source text in read-to-write tasks makes the tasks authentic and activates the writer's knowledge around a topic (Plakans, 2008; Weigle, 2004). These tasks have also been considered more fair for test-takers as they constitute a common information source for all of them (Weigle, *ibid.*). On the other hand, tasks with bare or framed prompts require background knowledge, depending on the subject matter. On the basis of this, the conclusion could be drawn that read-to-write tasks are generally more fair for test-takers and could lead to comparable results. However, this is not always the case and it depends on the testing situation, as the related literature indicates. The source text has been mentioned as making less proficient L2 learners rely heavily on the source text and develop or combine their ideas to a lesser extent (John & Mayes, 1990; Lewkowicz, 1997).

c. Task content

The content of a task comprises all those elements that are articulated in prompts. Table 2 presents a task analysis framework, which includes all the elements that make up task content. This framework has been adapted from the CEFR grid for task analysis and has been modified to serve the purpose of analyzing genre-based writing tasks¹⁸. It should be noted that task analysis frameworks may vary according to the nature and type of each writing task.

¹⁸ The term 'generic process' used in the framework comes from Knapp & Watkins (2005), who view genres as processes: to describe, to argue, to narrate, to explain and to instruct.

For assuring validity, item writers should follow test-specific guidelines and illustrative descriptors when designing the content of each task. What is more, to make the writing tasks as fair as possible, they should consider the target population. The two writing tasks below help illustrate these points further. These writing tasks come from different examination batteries.

Table 2: Task Analysis Framework for genre-based writing tasks

<u>TASK ANALYSIS FRAMEWORK</u>
➤ GENRE (TEXT TYPE)
➤ COMMUNICATIVE PURPOSE
➤ GENERIC PROCESS
➤ TOPIC OR THEME
➤ DISCOURSE ENVIRONMENT (CONTEXT)
➤ IMAGINED IDEAL WRITER AND AUDIENCE
➤ REGISTER / STYLE

Example 1

The writing task below comes from the KPG exams¹⁹. The initials KPG stand for Kratiko Pistopoiitiko Glossomatheias, which is the Greek State Certificate for language proficiency. The KPG exams address people living and studying or working in Greece. This is because one of the activities comprising the writing and speaking modules of the exam is a mediation activity where test-takers are expected to produce English output on the basis of Greek input. Test-takers' age varies depending on the level being examined. C1 level writing tasks, for example, address mainly adults and they are not designed for young test-takers. In KPG, writing tasks are designed according to CEFR guidelines and illustrative descriptors. Furthermore, it must be noted that writing task design and assessment follow a genre-based approach which considers that lexicogrammar becomes meaningful only when it is linked to text purpose and function (Mitsikopoulou & Dendrinou, in press).

The task below is an 'intralinguistic' written mediation C1 level task. Dendrinou (2006) has used the term intralinguistic mediation to differentiate it from interlinguistic mediation where the candidates are provided with a source text in their native language and are asked to produce

¹⁹ Readers can find KPG exams at the RCEL website (www.uoa.gr/english/rcel), as well as at the Ministry of Education site (www.kpg.ypepth.gr).

their output in the foreign language by selectively extracting the necessary information.

KPG – C1 LEVEL, MAY 2008

Based on the information from the website below, write a letter (180-200 words) to your favourite **12-year-old nephew** Ronnie, **explaining** why downloading copyrighted music is unethical and **trying to convince** him that he should stop doing it.



The screenshot shows a Windows Internet Explorer browser window. The address bar displays the URL: C:\Documents and Settings\USER\Desktop\May08\C1 exam_May08\C1_M2_May08\Downloading Copyrighted Music.htm. The page title is "Downloading Copyrighted Music". The website header includes the logo for "Connect with Kids" and navigation links: About Us | Products | Praise | Parents | Educators | Store | Contact Us. The main content area features a small image of a person at a computer, followed by the article title "Downloading Copyrighted Music" by Robert Seith, CWK Senior Producer. Below the title are links for "TipSheet", "What Parents Need to Know", and "Resources". A quote from Jonathan Morse, 18, is displayed: "I really think that the music companies are overreacting. I still buy the same number of CDs that I would before I learned about downloading music online." Below the quote, the text reads: "A couple of clicks, a few minutes to download and you've got a copyrighted music file - a song for free!" and "Pretty much everyone that I know downloads music on their computer," 18-year-old Jonathan Morse says.

Why is downloading music without paying for it an illegal act?

Music United for Strong Internet Copyright, a network of songwriters, musicians and performers dedicated to preventing the illegal reproduction of music, suggests discussing with your child the following reasons why he or she should not download free music:

- Stealing music is against the law
- Stealing music betrays the songwriters and recording artists who create it
- Stealing music stifles the careers of new artists and up-and-coming bands
- Stealing music threatens the livelihood of the thousands of working people employed in the music industry

What can I do to ensure that my child doesn't break the law?

Help your child stop breaking the law by illegally downloading music. Help him or her resist the urge to steal by following the strategies cited by the Smithsonian Children's Medical Centre:

- Teach your child about ownership at a young age
- Inform him or her about how s/he can have what s/he wants without stealing
- Be a good role model
- Develop an open relationship with your child
- Recognize honest behaviour

This task requires that candidates read and understand the specific type of multimodal source text and selectively extract the necessary information to produce a text of a different genre, register and style than the original, suitable for the context of situation, as shown from the task analysis below (see Table 3). To perform the task successfully, candidates **must** have:

- ❖ the **linguistic competence** to produce a letter to their nephew explaining why downloading music is unethical
- ❖ the **sociolinguistic competence** to create a meaningful letter relaying the information in the source text in a way that is appropriate for the context of the situation

❖ school, social and practical **literacy** to interact and mediate between the source and the target text

The task has been designed on the basis of test specifications and CEFR illustrative descriptors. In addition, the test scores demonstrated test construct validity. However, as can be inferred from the task analysis below (see Table 3), this is a demanding task which requires test-takers experienced in in this type of genre restructuring. In addition, test-takers are required to produce a letter to their nephew assuming the role of an adult. When administered, this task was not considered unfair because item writers designed it assuming that test-takers are young adults – the KPG exam is actually designed for test-takers who are 16+. If the target group had been younger test-takers, they might have been in a disadvantaged position over older ones if they had not received the necessary training. What is more, understanding and dealing with the source text which is a webpage text necessitates computer literacy on the part of test-takers. Those who are not familiar with this text type may find it difficult to interact with it as required. Finally, it is subject matter that might not be familiar or appeal to everybody.

Table 3: Task analysis

<u>Source text</u>	<u>Target text</u>
Genre (text type): Webpage text	Genre (text type): letter
Communicative purpose: inform	Com. purpose: explain & convince
Generic process: explain & instruct	Generic process: argue
Topic or theme: music downloading	Topic or theme: music downloading
Discourse environment: website	Discourse environment: personal domain
Imagined ideal writer: writer of the text	Imagined ideal writer: uncle
Imagined ideal audience: general public	Imagined ideal audience: nephew
Register / style: impersonal / neutral	Register / style: personal / informal

Example 2

The B2 level writing task below comes from Cambridge ESOL exams. This task requires that test-takers write a review of a play they have seen by describing and evaluating the characters, the costumes and the plot and also by explaining why they would recommend the play to others. What should be noted is that this task was not obligatory for test-takers as they were offered a choice²⁰. This task actually requires particular literacy skills. Test-takers who have not been to

²⁰ In the writing part of Cambridge ESOL, test-takers have to respond to two tasks; the first one is obligatory and the second one is

the theatre may not be able to produce a satisfactory output due to lack of background knowledge. Especially, since the prompts are not text-based, there is not any stimulus for test-takers to resort to. Also, the required text type (i.e. a review of a theatrical play) might not be familiar to everybody and test-takers who have not been exposed to such text types might not be able to respond to task requirements successfully.

You recently saw this notice in an English-language magazine called *Theatre World*.

Reviews needed!

Have you been to the theatre recently? If so, could you write us a review of the play you saw? Include information on the characters, costumes and story and say whether you would recommend the play to other people.

The best reviews will be published next month.

Write your **review**. (120-180 words)

Conclusion

From the above, it seems crucial, therefore, that among various parameters, item writers consider the task type and the task content in relation to the target population when designing writing tasks. As mentioned in CEFR (2001), for establishing the potential difficulty of a given task for a particular learner (test taker), item writers should consider: a. cognitive factors (task familiarity, required skills), b. affective factors, and c. linguistic factors. Attempting to design fair writing tasks is not an easy thing to do; it would seem impossible to make writing tasks cater for everybody. Nevertheless, by following test specifications and by considering the target population, test designers should be able to design tasks which are valid and as fair as possible.

References

- Bachman, L.F. & A.S. Palmer (1996) *Language Testing in Practice*. Oxford: OUP
- Council of Europe, Ed. (2001) *Common European Framework of references for languages: learning, teaching, assessment*. Cambridge: CUP
- Dendrinis, B. (2006) Mediation in Communication, Language Teaching and Testing. *Journal of Applied Linguistics* 22: 9-35
- Johns, A., & Mayes, P. (1990) An analysis of summary protocols of university ESL students. *Applied Linguistics* 11: 253-271
- Knapp, P. & Watkins, M. (2005) *Genre, Text, Grammar: Technologies for Teaching and*

offered as a choice among four tasks.

Assessing Writing. Sydney: UNSW Press

Kroll, B. & J. Reid. (1994) Guidelines for Designing Writing Prompts: Clarifications, Caveats, and Cautions. *Journal of Second Language Writing* 3(3): 231-255

Lewkowicz, J. (1997) Investigating authenticity in language testing. Unpublished doctoral dissertation, University of Lancaster

Mitsikopoulou, B. & B. Dendrinos (ed.). in press. *The KPG writing test in English: A Handbook*. University of Athens: RCEL Publications, Series editors B. Dendrinos & K. Karavas)

Plakans, L. (2008) Comparing composing processes in writing – only and reading – to-write test tasks. *Assessing Writing* 13: 111-129

Purves, A. (1992) Reflection on research and assessment in written composition. *Research in the Teaching of English* 26: 108-122

Weigle, S. C. (2004) Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing* 9: 27-55

Xi, X. (2010) How do we go about investigating test fairness? *Language Testing* 27(2): 147-170

voikonomid@gmail.com

Vasso Oikonomidou holds a B.A. from the Faculty of English Studies (University of Athens) with distinction, and an M.A. in Applied Linguistics (Reading University, UK). She has had EFL teaching experience since 1999. She is a PhD student and is working as a research assistant at the RCEL, University of Athens, Greece.

Using the CEFR in the foreign language classroom

Anne Dragemark Oscarson,

University of Gothenburg, Sweden

Mats Oscarson,

University of Gothenburg, Sweden



Introduction

This paper describes work undertaken in order to investigate the usefulness of CEFR-related concepts and materials in everyday classroom settings. The educational context is a research project which aimed to explore language teachers' acquaintance with and use of various forms of assessment, in particular approaches that support the development of autonomous study skills and learner self-assessment of achievement. Various methodological procedures were employed for the collection of research data, including a questionnaire survey, teacher interviews, and an empirical try-out.

For quite some time in Sweden there have been reports of grading practice not being comparable or "fair", i.e. in the sense of not being performed in the same manner throughout the country (Skolverket 2004a, 2004b, 2006; Tholin, 2006). One of the reasons behind a new curriculum to be introduced in 2011 is in fact the need to make standards of grading more equivalent between schools; the previous curriculum goals and grading criteria were considered too open to interpretation.

As grading is a difficult task, and as teachers have many different aspects to consider when deciding on their marks, there is always a risk that they use more tests than necessary - not always fully considering what the tests employed actually measure. It is of course unfortunate for our students if their grades are not given in accordance with official guidelines in the curriculum as their whole future may in fact be determined by their school record. In Swedish research conducted by Selghed (2004), for example, it was noted that there are teachers who still grade according to an older, now abandoned, relative grading system and who use predominantly quantitative rather than qualitative achievement criteria when grading.

In a criterion-related grading system it is very important for the students to understand what the goals are according to the curriculum, and what is expected of them. They need to know what

characterises any given level of attainment and what is required for a certain grade. Through other research projects we have conducted we know that it is a common opinion among students that teachers do not always see what they can actually do with the language. It has also been argued, as a possible added complication, that tests are sometimes used as a means to “discipline” students rather than as a means to further and assess their learning (Lynch, 2001; Shohamy, 2000). Furthermore, research has shown that students who feel that they are more actively engaged in the process of learning and evaluation of results, through self-assessment practices for example, tend to be better learners (Dragemark Oscarson, 2009).

The Project

The project, *The Teacher’s Extended Assessment Role*, financed by the Swedish Research Council, had three parts. One was a nationally representative questionnaire survey of teachers’ assessment practices. Another part was an interview study based on personal construct psychology (Kelly, 1963) and using so-called repertory grid elicitation technique. For further information on this part of the project, see Apelgren (2010) and Oscarson & Apelgren (2011). Finally, there was a small empirical study where teachers were requested to use so-called alternative methods of assessment in their ordinary classroom teaching in order to estimate their usefulness.

The Questionnaire Survey

The aim of the survey was to investigate the basis on which teachers determine students’ grades in the foreign language classroom. Another aim was to identify the needs for further training and information that teachers experience in the area of language assessment and student grading.

The survey involved a nationally representative sample of 605 language teachers distributed over some hundred lower and upper secondary schools. The age groups taught were primarily 15-18 year-olds and the languages were mainly English, French, German, Italian, and Spanish. Basically the questionnaire requested teachers to respond to items which related to their task as evaluators of their students’ learning. The focus was on principles of grading, and on their personal assessment preferences and practices.

One of the key items in the questionnaire was:

“What types of evidence and assessment instruments do you take into account and use when

you assess/grade your students? Tick each option which is applicable in your case.”

The response options covered all the different sources of information that Swedish language teachers tend to use and which reflect variable attitudes to student assessment.

Three sources dominated among the answers:

- classroom observation of free oral communication
- own, teacher-produced tests
- essays on given topics

There was practically no difference in preference between lower and upper secondary school teachers. On average, each of the three sources were reported to be used by about 95% of teachers in both school forms.

Other important assessment sources referred to were:

- standardized (national) Tests
- classroom observation: oral communication on a given topic
- pre-announced homework tests

The first two of these were somewhat more commonly employed in the upper secondary school. Pre-announced homework tests tended to be more frequently used in the lower secondary school.

At the other end of the scale we found that the least common assessment approaches were such that involve active engagement by the students themselves in the assessment process:

- alternative Assessment in the form of portfolios
- alternative Assessment in the form of peer-assessment
- student-produced tests

On average, 14% of the total sample of teachers indicated that they used these or similar forms of help when they assess the results of their instruction and the students' work. The picture was about the same when the two levels were compared: both lower and upper secondary school teachers placed the above techniques at the *end* of their lists of adequate assessment sources.

The response patterns are likewise similar across language groups, that is, between English as

a Foreign Language and other foreign languages. An hypothesis that over-all assessment strategies vary significantly depending on level and language taught was thus not confirmed.

The outcome of the survey made it clear that so-called “alternative” assessment forms, including the use of the European Language Portfolio (henceforth ELP) and CEFR-based assessments, are *not* employed very often and this was also generally confirmed in the subsequent interviews with a sub-sample of teachers (Oscarson & Apelgren, 2011). More traditional methods play a much more important role, especially for grading purposes. In this respect, there is little difference between teachers of English and teachers of other languages.

Other factors than traditional linguistic ones also proved to play a distinct role in assessments. Class attendance and motivation (interest, dedication etc) were also rated as important variables in teachers’ final grading of students, as the figure shows.

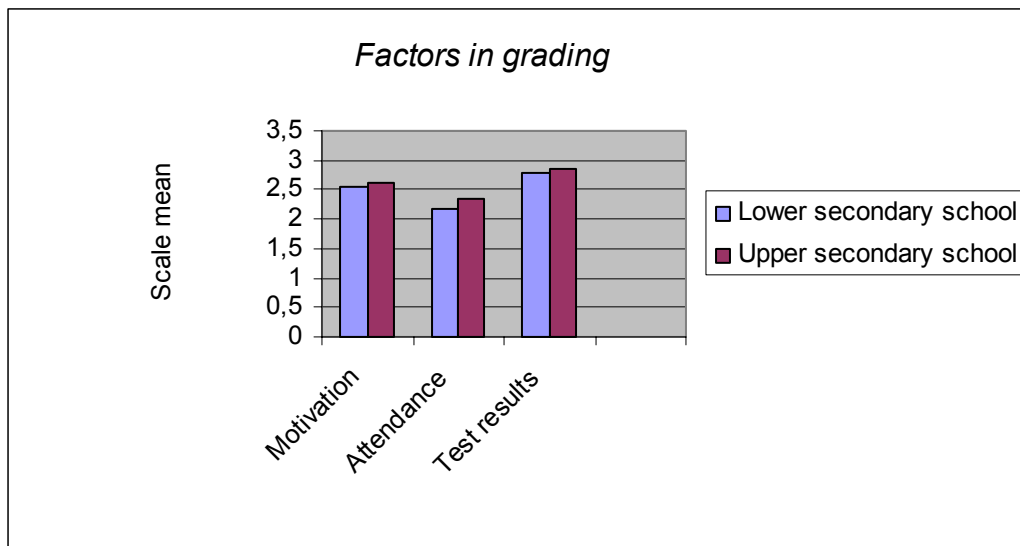


Fig. A: comparison of three factors underlying grading in Swedish FL classes

Attendance and (in particular) motivation tend to carry almost the same weight as test results in the overall grading of students’ performance.

The Empirical Study

This part of the study was devoted to exploring the knowledge and actual use of some alternative forms of assessment such as self-assessment, peer assessment, and portfolio assessment. Their application is based on the theory that metacognitive skills such as self-regulation and self-monitoring are important for the development of students’ learning (Hartman, 2001). Attaining such skills is considered important in Swedish curricula.

Seven teachers were selected on the basis of their declared interest in participation in the project. They taught either French, Spanish and/or English. Without previous experience, they tried out different forms of self- and peer assessment with Grade 6-12 students during one school term. They could choose the form of assessment themselves and most of them opted for the global CEFR scales or the Swedish version of the ELP (age bracket 12–16) with its self-assessment grids. Two of the teachers wanted to use DIALANG but had some difficulty with the technology so that they partly had to give up the idea. The results presented here are from interviews and reports from four of the teachers who were working at three different lower secondary schools and who all used the global CEFR scales alone or as a part of the ELP with their students.

April, teacher of Spanish (Grade 9)

April taught Spanish in a grade 9 class and had found that her students had difficulty understanding the language, even though they had studied it for three years. She had previously filmed her students at the beginning of the school year and then towards the end, so that they could see their own progress. During the project, April used the self-assessment grid in the ELP as a tool in a particular learning sequence. The students listened to an episode in a Spanish TV series, practised retelling a part of the episode, and then assessed their own ability to communicate their assigned task to their peers using the appropriate CEFR scales.

April's students were quite enthusiastic about being able to assess themselves according to the CEFR, but April felt that she should have adjusted the descriptors somewhat to fit the actual lesson content better. The students had a hard time generalizing. For instance, if the relevant scale said that a person at a particular level was able to talk about his or her family and the episode did not concern a family, the student tended not to know how to go on. April also found that she should have started with her beginners, i.e. a grade 6 class and not with students in the last year of lower secondary school (grade 9). The ninth graders were very aware of the Swedish course syllabus goals and the grading criteria, and this caused a certain degree of uncertainty, as they wanted to compare and relate the CEFR scale descriptors to the criteria they were used to. Having two sets of scales to relate their achievement levels to confused them. In the follow-up interview April said that she wished that the CEFR scales could be systematically linked to the levels in the Swedish school system, or vice versa, but she also said that the best thing about using the ELP was that the students could see their own progress in a

way that the Swedish syllabus goals and grading criteria did not allow for.

**Kathy and Ursula, teachers of Spanish and English
(Grade 6 French, Grade 7 Spanish and Grade 9 English)**

Kathy and Ursula worked at the same school. Both taught grade 9 English. Kathy also taught a grade 7 Spanish class, where students had studied the language for two years, and Ursula a beginners class of French, i.e. grade 6. Generally they found it more difficult to motivate their students to study Spanish and French than English.

They started the term by having all their students compare their own perceived language levels with the global CEFR scales. Compared with their own assessment of their students, Kathy and Ursula found that the students over-assessed their language levels in English but that their assessments of their levels of French and Spanish were quite accurate. The students then worked independently and/or in groups on areas that they felt they needed to improve on, based on their own CEFR assessments. For a duration of two weeks students worked with, for example, different writing tasks or speaking exercises and/or points of grammar. Following this period the teachers presented their students with the ELP and let them fill in the self-assessment grids, the language passport and the language biography, partly as a way of motivating the students for further language learning. In this way Kathy and Ursula trained their students in self-assessment throughout the term and some of the students were able to use DIALANG, but due to technical problems not all of them could do so. Towards the end of the term the students then self-assessed their language levels in a conference with their teacher using both the Swedish grading criteria and the CEFR.

The students' evaluation showed that they appreciated the opportunity to choose the language learning focus themselves from their own perceived needs, and that they enjoyed working with the ELP. Both teachers experienced that their students made relevant choices of learning tasks to improve their language but that this took more planning on their own part than they had expected. Working independently, oral production was the most problematic area and, in retrospect, the two teachers involved realized that the students should have been encouraged to film or record their work themselves. A general language portfolio could also have been of great help in documenting their work, as would have more frequent teacher/student conferences. According to these teachers, awareness raising work like this should start early – in grade 6 or grade 7 at the latest. At the end of the project the two teachers observed that the self-

assessments their students made of their language levels, using the CEFR, were generally realistic.

Sarah, teacher of French (Grade 7)

Sarah taught French in grade 7. She started by making a simple inventory in her class on why they wanted to learn French. 14 of the 19 students said that they wanted to learn to *speak* the language. In response to this, Sarah focused on oral productive skills during the whole term.

At the start of the project Sarah let her students self-assess their speaking ability using the ELP for the early ages but soon realized that she should have used the one for the appropriate age group, as most of the students found the activities a bit childish. Her students marked their own language levels as quite high when it came to listening and speaking. Towards the end of the term, after a lesson that focused on reading, understanding and then retelling a French text in smaller groups, the students assessed their speaking proficiency according to the CEFR.

Sarah described how the students had difficulties in interpreting the descriptors at first, but after discussing them in class together they became, according to her, quite accurate in their assessments of their oral production, which mostly landed between A1 and A2. An example of how the students reasoned was when they said that it was easy to answer personal questions in French and to speak about their family, but that they had difficulties in posing questions themselves. Sarah's conclusions were that she should start using the ELP earlier, at the beginning of the course in grade 6, as students were not ready to assess their language level on their own without help and they obviously needed training.

Summary and conclusions

In conclusion, results of classroom observation, written assignments such as essays, and conventional language tests are the criteria primarily relied on when Swedish foreign language teachers assess their students' levels of learning. Two methods used for information gathering, a questionnaire and interviews, produced very similar results. Furthermore teachers use much the same sorts of assessment sources at different levels of teaching as well as in the different languages taught. In addition to observed performance and test results, non-subject-matter variables (class attendance, motivation) play a very important role in teachers' grading of student achievement.

The empirical study showed that the four teachers in the project who specifically used the CEFR global scales, either by themselves or as part of the ELP, in their classroom work with their students, experienced that their students developed an increased awareness of the curricular goals and that they themselves developed an increased awareness of their own teaching practice. They further reported that their use of the CEFR enhanced not only language awareness but also motivation, and also helped to make assessments more transparent. The problems they faced concerned practicalities and the fact that the CEFR levels and the national grades used in the Swedish school system are not directly aligned or comparable.

The four teachers also expressed the need for students to practise self- and peer assessment from an early stage in language learning. They reported a need to provide such practice on a continuous basis and that the ELP is one helpful tool to achieve this. As a result the teachers expressed their intention to continue working in this manner.

The implications of the study are that teachers, at least in Sweden, need in-service-training as well as help in increasing their competence in other forms of assessment than the more traditional summative ones. The global CEFR scales and the ELP can be valuable tools to help students develop heightened language learning awareness – and thereby to improve their study results.

References

- Apelgren, Britt Marie (2010) *Construing Learning and Assessment in the Foreign Language Classroom: the teacher as the meaning-maker*. In *Construing PCP: New Contexts and Perspectives: 9th EPCA Conference Proceedings*. (Eds.) Dorota Bourne and Martin Fromm. Norderstedt: Books on Demand.
- Dragemark Oscarson, Anne (2009) *Self-Assessment of Writing in Learning English as a Foreign Language. A Study at the Upper Secondary School Level*. PhD Dissertation. University of Gothenburg: Acta Universitatis Gothoburgensis
- Hartman, Hope J. (2001) *Metacognition in Learning and Instruction. Theory, Research and Practice*. Dordrecht: Klüwer
- Kelly, George (1963) *A Theory of Personality*. USA: Norton.
- Lynch, Brian K. (2001) Rethinking assessment from a critical perspective. *Language Testing*, 18, (4), 351-372
- Oscarson, Mats & Apelgren, Britt Marie (2011) Mapping language teachers' conceptions of student assessment procedures in relation to grading: A two-stage empirical inquiry. *System*, 39, 1
- Selghed, Bengt (2004) *Ännu icke godkänt. Lärares sätt att erfara betygssystemet och dess tillämpning i yrkesutövningen*. PhD diss.. Malmö: Malmö högskola, Lärarutbildningen
- Shohamy, Elana (2000) Fairness in language testing. In Kunnan, A.J. *Fairness and validation in language assessment*. Cambridge: Cambridge University Press

Skolverket (2004) *Grundskolan i blickpunkten. Sammanfattning och slutsatser från Nationella Utvärderingen av grundskolan 2003*. Stockholm: Skolverket

Skolverket (2006) *Lusten och möjligheten. Om lärarens betydelse, arbetssituation och förutsättningar*. Rapport 282. Stockholm: Skolverket

Tholin, Jörgen (2006) *Att kunna klara sig i ökänd natur. En studie av betyg och betygskriterier – historiska betingelser och implementering av ett nytt system*. Doctoral Dissertation. Borås: Högskolan i Borås.

Anne.Dragemark@ped.gu.se#

Mats.Oscarson@ped.gu.se

Dragemark Oscarson, Anne PhD, Senior Lecturer, University of Gothenburg. Research interests include language education and assessment of language skills, especially classroom and self-assessment. Has previously worked with test development and been involved in several Swedish and European language projects. Presently engaged in teacher training at the Department of Pedagogical, Curricular and Professional Studies.

Oscarson, Mats PhD, Professor emeritus (Education), Department of Pedagogical, Curricular and Professional Studies, University of Gothenburg. Fields of interest include FL language instruction, testing and assessment (notably student self-assessment) and teacher training. Has directed several research projects financed by the Swedish Research Council and the Swedish Agency for Education and has participated in international research initiated by the Council of Europe and the EU.

What to teach and assess from A1 to C1

Susan Sheehan

British Council, UK



Introduction

The British Council and EAQUALS have joined to create the Core Inventory for General English. The Core Inventory is designed for adult learners on general English courses. It includes grammar, vocabulary, functions and notions, discourse markers, scenarios and exponents. The intention of this project was to make the CEFR accessible to teachers and adult learners of General English. It was an attempt to answer the question put by teachers of what the CEFR means in terms of classroom activity. The project had two further aims, to make the teaching/planning process more transparent to learners by providing clear learning objectives and to provide support for self-directed study by providing a guide to essential language for study. In the paper below I outline the five stages of development, the sources the project drew on, suggestions on how to use the inventory and the development and purpose of the scenarios.

5 stages of development

The Core Inventory was developed through an iterative and collaborative process. At a series of workshops, experienced and expert practitioners commented on work completed to date and offered suggestions for the subsequent stages. These practitioners were drawn from the two partner organisations and examination boards. The five stages were as follows:

- data collection and analysis
- creation of the Inventory
- writing the exponents
- identifying text types
- writing CEFR-based scenarios

As discussed below, a number of different data sources were drawn on. The data were analysed to find consensus. Points which were common to 80% of the data sources were defined as “core”. This led to the creation of the first version of the Inventory. Examination boards then provided input on the language points they considered relevant. Their input at the higher levels

was particularly valuable. At this stage the Inventory was finalised. Some language points which did not reach the 80% criteria but were considered, nevertheless, as important by practitioners were included in the Core Inventory. They were labelled as “less core” to differentiate them from the “core” items. The next stage was to write sentences to exemplify the language points, i.e. exponents. The next stage was to conduct an analysis of CEFR descriptors to identify source texts for different CEFR levels. The final stage was to write the scenarios which will be discussed in detail below. The project was completed in one year.

Sources

The project drew on four main sources. We were interested in establishing the practitioner’s point of view so we consulted sources to provide insights into this perspective. Learner language databases and corpora were not included as data sources, as other projects, such as the English Profile project are already conducting valuable work in this area. The main data sources drawn on were the following:

- an analysis of the language implied by CEFR descriptors;
- an analysis of content common to the syllabuses of EAQUALS members whose CEFR implementation was a point of excellence;
- an analysis of content of different series of popular coursebooks;
- teacher surveys.

The highest levels of consensus were found at B1. This may reflect the influence of the Threshold Level, published in 1976. Good levels of consensus were found at A1 to B2. At C1 context and learning purpose seemed to dictate the content of the coursebooks and school syllabuses and so the level of consensus was reduced. We were unable to establish any significant consensus at C2 and for this reason it was decided that the Inventory would not include this level. The teacher surveys showed that the study of grammar is important at A1 to B2 but becomes less important after that. Lexis becomes increasingly important from B2 on.

How to use and not to use the Core Inventory

The Core Inventory is intended for use as a reference work. As the name suggests, it is the core, it is not the whole. Teachers and syllabus writers will define the total content of a course. The Inventory provides guidance and support for those who are involved in course design. It provides the foundation for courses for institutions which aspire to reflect the aims of the CEFR in their course aims. The Inventory documents one approach to realising an “action-orientated” approach to language learning and language use described in the CEFR. It is only one possible

approach to achieving such an aim and institutions could adapt our methodology to make an inventory for their own situation.

Teachers will continue to negotiate course content with learners and to conduct needs analysis. The Inventory is a tool which can help the negotiation process and provide a common language for such discussions. Learners will continue to have specific reasons for learning English and will continue to look to their teachers for guidance and support as to how best to achieve their learning goals. So, every classroom will continue to be a unique environment which reflects the needs and knowledge of those contained within it. The Inventory is in no way a substitute for this negotiation process and does not replace listening to students.

We have documented examples of good practice and offer these to the wider audience of English language practitioners. Those practitioners will then decide on how to make the Inventory relevant to their situations. They may, of course, decide that the relevance is limited and so choose not to use it. The Inventory places each language point in the place where it is of most relevance to the classroom. It does not contain information about when and how language points should be recycled. These decisions are left to syllabus writers and teachers. The Inventory does not offer guidance on how language points should be introduced, practised or developed so the teacher will make such decisions, as they always have done, taking into account the local context and the needs of the students in any particular group. The Inventory does not state when learners will have mastery of a language point.

Scenarios

The Core Inventory includes illustrative scenarios for levels A1 to C1. A scenario starts with a real-world situation such as a business meeting. The domain, context, tasks, activities and texts for the situation are defined. These variables are taken from the CEFR and reflect its “action-orientated approach”. To these, “can-do” descriptors, quality criteria and aspects of competence (e.g. strategic, pragmatic and linguistic) are added. The “Can-do” descriptors function as objectives. The quality criteria are there for evaluation. The aspects of competence are enabling objectives.

The scenario shifts into the classroom through the scenario implementation which outlines teaching/learning and assessment activities related to the competences needed to perform the real world tasks. The number of activities included in different scenarios varies depending on the type of scenario. Some include only assessment tasks, others a mixture of assessment and

teaching tasks and some include only teaching activities. This demonstrates the flexibility and of the scenario model and its applicability to a variety of teaching and assessment activity types. Assessment in this case is not restricted to the teacher-led type. Peer and self-assessment are also included in the scenario model. During the writing workshop the authors often found it difficult to complete the first section of the scenario as we tended to think in terms of classroom activities rather than real world situations. Teachers are encouraged to work with colleagues and students to create scenarios which are relevant to the needs and aspirations of students studying at their own institutions. The scenarios provided in the Inventory show only a few of the many variations possible.

Conclusion

The Core Inventory for General English will help to make the CEFR tangible and provide support and guidance for teachers and syllabus designers. It will aid learners to make the connections between classroom activities and real world needs. It does not tell people what to teach. In addition to the Core Inventory described in this paper the Core Inventory Project has two other products. One is The Essential Guide which contains only “core” functions, grammar, lexis, and discourse markers together with a brief summary of the project aims and guidance for use. The second product is a set of classroom posters. These include “Can-do” descriptors, core language points with exponents and qualitative criteria. The posters are designed to make the content of the Inventory easily available to both teachers and learners and provide a focus point for classroom discussions on course content and planning.

The Core Inventory is based on consensus and good practice from expert and experienced practitioners. We believe we have held a mirror to the profession and recorded what is being taught in classrooms. We did not aim to be prescriptive. Nor did we aim to be totally comprehensive. We have produced a resource which we believe will help learners and teachers adopt an “action-orientated approach” to language learning and teaching.

The Core Inventory for General English is available for free from <http://www.teachingenglish.org.uk> and from www.eaquals.org.

Susan Sheehan works for the English Language Innovations team at the British Council. She is Adviser Learning and Teaching. Her areas of specialism are testing and the CEFR. Susan has delivered courses on testing and assessment in many countries and managed the development of the new British Council placement test.

Putting the CEFR to Good Use: Activities and Outcomes in Finland

Sauli Takala, Professor (emeritus)

Center for Applied Language Studies, University of Jyväskylä, Finland



1. Introduction

I will present a selective account of the activities on and with the Common European Framework of Reference for Languages (CEFR) in Finland. This will also reflect my own perspective, drawing on long contact and association with the Council of Europe's modern language project.²¹ This gave me a good opportunity to help mediate the CoE initiatives to the development of language education in Finland as I was a regular member of several subsequent curriculum development teams in modern languages. Finland tended to be among the "early adopters" of the many CoE contributions to the updating of language teaching and learning (Takala 2006). The curriculum of the new comprehensive-type basic school in 1970 reflected the recommendations of the Ostia and Ankara conferences in 1966, which featured the "four skills", becoming acquainted with the target language culture and developing a positive attitude to its speakers. In the mid-1970s the functional-notional approach embodied in the Threshold level was adapted to school use. Subsequent curricula strengthened the role of learner autonomy and responsibility, self-assessment and reflection and cross-cultural competence.

Why did Finnish language educators and decision makers have such a favourable attitude to the approaches and tools developed under the umbrella of the CoE modern language project?

There is no research information on this but I will present some personal views. Perhaps the most important reason is the fact the CoE developed a coherent and far-sighted general policy for its work in promoting broad-based European cooperation in developing new initiatives in language education. It was able to draw on top experts from a variety of countries ensuring that

²¹ My first brush with the CoE was in 1968 when I attended its seminar on language testing/assessment in Skepparholmen, Sweden. In 1976 I attended the symposium in Holte, Denmark, whose theme was modern languages in primary education. A more active role occurred at "a meeting of experts on the extension of the threshold to school education" in Strasbourg in 1976. In the 1990's, I was a member of the advisory group related to the development of the CEFR and in the 2000s a member of the working group developing the manual for relating examinations to the CEFR.

the work reflected the broad range of priorities, preferences and educational cultures. The experts' preparatory work represented cutting-edge knowhow and solid experience in language education. The project was managed effectively and skillfully and displayed great cross-cultural awareness and sensitivity. There was never any attempt to impose any ideas, as all work was based on extensive consultation, interaction and feedback. Through publications and numerous seminars many language educators got acquainted with colleagues in other countries and learned from each other and acquired a better shared metalanguage. For me and surely for many others, involvement in the CoE modern language project activities has been an "invisible college", providing learning experiences that one would not get from "ordinary" academic experience.

2. Some notes on the development of the Framework

The development of the CEFR was a long process, initiated at an intergovernmental conference in Rüschiikon, Switzerland, in November 1991 together with the work on the language portfolio. The development work was part of the medium-term project "Language Learning for European Citizenship" (1990-1997). A working party was set up (D. Coste, B. North, J. Sheils and J.L.M. Trim) to produce a draft framework. An advisory group was set up in 1993 and I was appointed one of its members. In 1995 we had Draft 1 to review, and it was made available in December 1995.²²

Throughout the different drafts the basic principles of flexibility, openness, dynamism and non-dogmatic stance were adhered to as well the rejection of prescriptivism and the acceptance of its evolution (Trim 2007). There were, however, some changes made. It was expanded in the process: Draft 1 contained 204 A4-pages, Draft 2 213 A4-pages and the printed book 257 pages. Some visual illustrations were removed and I am afraid that we, the advisory group, provided unhelpful advice on this score, as the illustrations were quite useful. The scales were presented in the first two drafts in appendices, and one might argue that there was something of an apologetic tone in introducing them. There was a proposal in the advisory group to remove them entirely but it was decisively turned down by the majority. In the event, they were brought forward and incorporated in the body of the text. There was also some other reorganization of the content and, for instance, the currently strongly emphasized notion of "plurilingualism" in an

²² It was entitled "Common European Framework for language learning and teaching. Draft 1 of a Framework proposal". The second draft was entitled "Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference (published in March 1997 and sent in 1998 for extensive field consultation). The present title was decided on when the framework was to go to the printers.

elaborated form was introduced first in the printed book.

It would be interesting to see a more in-depth analysis of the different versions, combined with the information provided by the authors in different contexts (see e.g. Coste 2006)

3. CEFR in Finland

In this article, I will include work done in Finland, both as a national and international activity as they happen to be closely intertwined. Takala and Kaftandjieva (2004) also provides an account of some of the activities.

3.1. DIALANG

It all started with the EU project DIALANG, which my home department, Center for Applied Language Studies, University of Jyväskylä, coordinated during the first phase (autumn 1996 - November 1999). During the early part of 1996 the idea of producing language tests took form and this activity was originally to be a pilot project like a number of other assessment projects within the EU. The project was, however, soon transferred to DG XXII/LINGUA. In this context the original idea of accreditation and the personal skills card was abandoned in favour of a diagnostically oriented assessment system, which was approved by the SOCRATES committee.

DIALANG was in many ways a novel approach to language testing and assessment. It developed a transnational assessment system with a large range of languages covered. It is diagnostically oriented, with one purpose to promote diversified language learning in Europe. It combines self-assessment and external assessment. It uses the Internet as the delivery system and reports the results in terms of the Council of Europe proficiency scale. This linking was decided as the use of the scales was seen to promote comparability across languages. I played an active role in developing the blueprint and I recall that I saw in the system an opportunity to “democratise” testing/assessment by trying to put the user “in the driver’s seat” and by serving his/her individual interests (to be “at his/her beck and call”) In this DIALANG displayed a similar sense of mission as EALTA (see also Erickson in this publication).

In practice, the CEFR Draft 2 self-assessment scales and communicative activity scales were used/adapted. We also reviewed and utilised the objectives definitions in Waystage, Threshold and Vantage. While we found them useful for test specification, we also noted that there was quite a lot of overlap in them and thus progression was not always very clear-cut.

DIALANG faced many daunting challenges: how to write test specifications, self-assessment statements, feedback statements and relate all this to the CEFR (see Alderson 2005). Of course, relating the outcomes to the CEFR was a huge challenge (Kaftandjieva, Verhelst & Takala 1999) and it became a “hot” topic after the CEFR had been published in 2001. It needs to be pointed out that standard setting required a new approach: from the usual task of setting one cut-score (failing/passing the standard), as many as five cut-scores were needed. This was done using as a starting point the “modified Angoff” method.²³

The results of a validation study (Kaftandjieva & Takala 2002), which was designed and conducted as a part of a pilot study of a standard-setting procedure specifically designed for the purposes of DIALANG, provided strong support for the validity of the CoE scales for listening, reading and writing. These findings not only confirmed that the DIALANG assessment system was based on solid ground but they also had a broader impact, supporting the view that any further development of the CEFR could be undertaken on a sound basis.

3.2. Scale development

While the CEFR scales have become the benchmark in Europe and beyond, there were many scales developed and used before the CEFR. Indeed, Brian North (1995) reports in his PhD thesis that almost 30 scales (and about 1000 descriptors) were used in the Swiss project that led to the CEFR scales.

In Finland, the need for a national certificates system (YKI) was discussed in the early 1990s and introduced by an Act in the mid-1990s. A number of reasons were presented for the system, including the opportunity for adults to have a reliable assessment of their language proficiency irrespective of how they had acquired the skills and the possibility of using the data for

²³ Actually three different modifications of the modified two-choice Angoff method as well as three different modifications of the contrasting group-method were applied to the standard setting procedure. Multiple matrix sampling with incomplete equal-sized linked design was used to pilot the items. Item response theory was applied to item calibration. The One Parameter Logistic Model (OPLM) was chosen, because it combines the desirable statistical characteristics of the Rasch model with the attractive features of the two-parameter logistic model. Moreover, the OPLM computer program allows application of incomplete test design, which at that time was not possible with most of the other computer programs that applied the IRT approach to test development and analysis. The adaptive test construction design was based on the two-stage multilevel adaptive testing approach. The role of the routing test (pre-estimation) is played by the Vocabulary Size Placement Test and the self-assessment tools. The second-stage language test has three overlapping levels of difficulty.

assessing the overall national language proficiency level and for developing language education. The development of the system drew heavily on the Waystage, Threshold and Vantage specifications and adapted the 9-point scale of the English Speaking Union.²⁴ As the CoE 6-point scale entered the scene and gained growing attention and acceptance, a project was set up to calibrate the original scale to a new 6-point scale. This required considerable conceptual and empirical work and the new scale was successfully validated against the CEFR scale (Kaftandjieva & Takala, 2003).

Another large-scale project (2000-2001) was carried out to develop a system for harmonizing high-stakes assessment of compulsory language requirements in polytechnics (tertiary level). Scales were developed on the basis of the CoE 6-point scales but adapting them to make them more relevant for a LSP context. Their validation procedure was largely similar to the earlier work developed by Dr. Kaftandjieva. This and the building of an item bank (a new feature) is reported in an unpublished manuscript (Kaftandjieva 2001).

A third project is presented to show a different context of scale development. The current syllabuses for the basic school and the upper secondary school (2004 and 2003, respectively) continued the long-established orientation of communicative language teaching and cross-cultural communication but introduced as a new element target levels for grades 6, 9 and 12 using school-adapted CoE scales. The most important deviation is the introduction of three sub-levels at A1: A1.1, A1.2 and A1.3. There are two main reasons for this: qualitatively clear progress is the most rapid at the beginning stage and more fine-grain levels are needed for reporting progress. It is surely very demotivating if a pupil feels that he/she is making progress but is reported to be at level A1 for a long period of study. Another adaptation is that there is more attention given to constraints than in the CEFR where the descriptors are predominantly couched in positive terms without indicating any constraints. It was felt that the spelling out of constraints was useful for the purposes of assessment and grading. Teachers have tended to agree that this is a useful addition. The scales were subjected to a small-scale validation (Hildén & Takala, 2007).

3.3. Assessment of learning outcomes in the school system

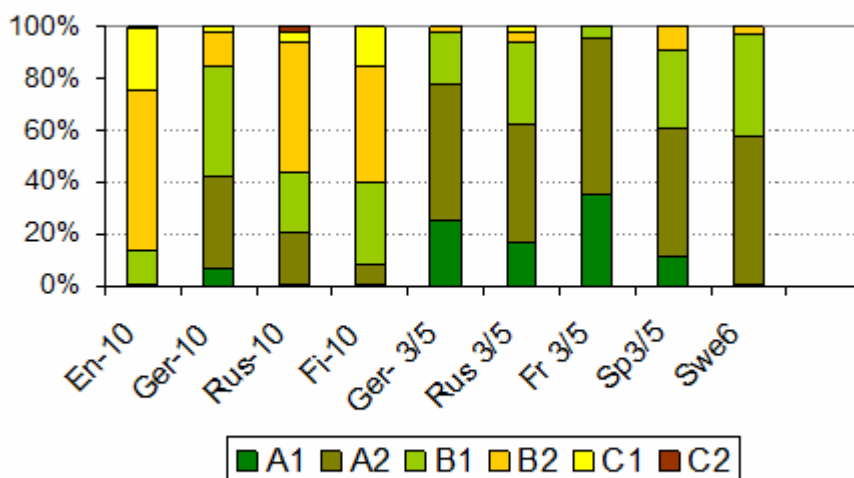
In the above, language proficiency was assessed in the DIALANG and AMKKIA projects. The “clients” of these were not “representative” of the ordinary school population. There have been a

²⁴ It is likely that the experience in developing and administering a language examination in several languages, drawing on the CoE tools was one of the main reasons why the Center for Applied Language Studies was encouraged to submit a proposal that led to the launching of DIALANG.

few studies where nationally representative studies have explored what level is achieved in language studies in the Finnish school system.

The first of these was a study of the level of achievement in English in the Matriculation Examination after 10 years of English (about 850 lessons, some 625 “clock” hours). The study (Kaftandjieva & Takala, 2002) was presented at a CoE seminar in Helsinki in the summer of 2002, which launched the process leading to the CoE Manual for relating examinations to the CEFR. Using basically the approach used by the authors in earlier studies it was established that about 60% of students had reached level B2, about 15% C1, 1-2% level C2 and about 15% level B1, which was the pass level. Figure 1 illustrates the levels reached in different languages with a 10-year course of study as against a 3-5 course study (started at the age of 13-14 or 15-16), and a 6-year course for Swedish (a compulsory language for Finnish-speaking students).

Distribution of Levels (%) in the Matric Exam (19yrs)



10: 10 years of study; **3/5:** 3-5 years of study; **6** - 6 years of study

Fig 1. Distribution of levels reached in the Finnish school system.

A few observations are worth pointing out. The level reached in English is much higher than in the other “long” languages (the same number of lessons). The good level of achievement in Russian and Finnish can be explained by the relatively large number of students who are strongly bilingual. This illustrates the fact that, especially for English, a substantial part of the level of achievement is explained by out-of-school use of English (“informal learning”). There is, in fact, a saying in Finnish that “English sticks to your clothes” – it is ubiquitous. It is sometimes

also called “the third national language”, with Finnish/Swedish being “the other national language”, respectively. The figure shows further that the level attained in the shorter courses is considerably lower than in the long courses. It also shows that learning outcomes in Swedish are quite low, mainly due to problems of motivation. It is also worth noting that the linkage of English and Swedish is more reliable than in the case of the other languages, which are presented here as tentative linkages.

There have been three national assessments with representative samples of basic school 9th graders. Fig. 2 shows the levels reached at the end of basic school after seven years of English (some 600 lessons, 450 “clock hours”). The results are reported by Tuokko (PhD thesis 2007).

Level in English (%): grade 9 (15-16 years; 7 years of English, Tuokko, 2007)

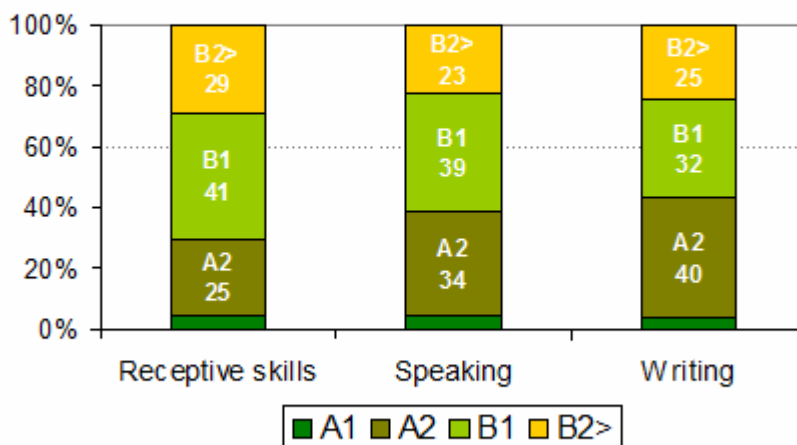


Fig. 2 Distribution of levels in English at the end of basic school

All four skills were assessed, but for the purpose of standard setting the receptive skills (listening and reading) were merged to reach a satisfactory level of reliability (as the relatively short tests did not possess a sufficient level of reliability). As the figure shows, the most common level reached was B1 (it was B2 at the end of the upper secondary school; cf. above).

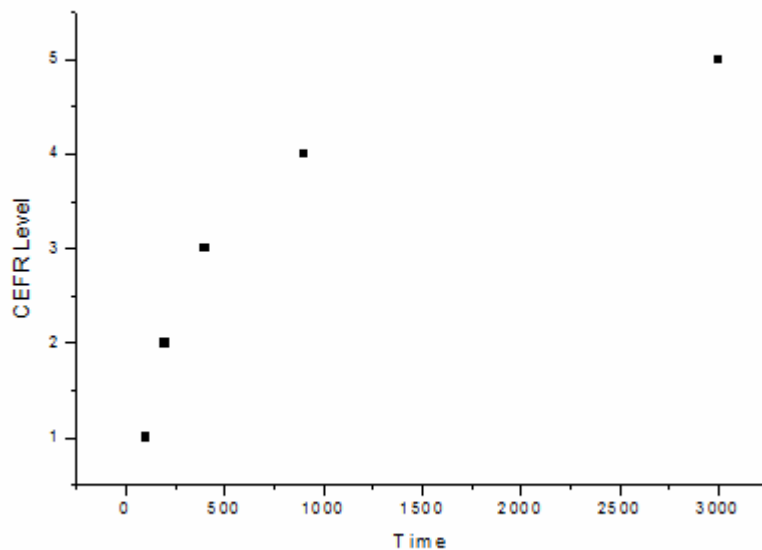
The level reached in Swedish (Finnish-speaking students) and Finnish (Swedish-speaking pupils) was lower. The level in Swedish (three years of study) was about A1.3 - A2.1 (Tuokko 2008). The level in Finnish was obviously higher (seven years of study) – A2.2 on the average but clearly lower than in the case of English (Toropainen 2010). For those students who studied

according to a more mother-tongue resembling syllabus (more or less bilingual students) the level was (obviously) higher – B1.2 on the average. An outcome which is worrisome in a school system which is strongly built on the premise of educational equality is the fact that the general level of attainment in Finnish was considerably lower in the coastal area in the mid-west than in the more bilingual-influenced southern coastal area.

4. Conclusion and discussion

Overall, it is probably fair to say that the rather extensive work on and with the CEFR in Finland has been a positive and rewarding experience and that it has been quite successful. On the other hand, the implementation of the CEFR applications in schools and classrooms (eg. in assessment/grading/examinations) has been slow and not very systematic.

One of the conclusions reached during the work with and on the CEFR is that there is qualitatively fast progress in the lower stages of language proficiency. After this it takes increasingly more time (exposure, use) to reach subsequent levels. This is illustrated tentatively in Fig. 3 (Level 1=A1, 2=A2 etc). The time scale represents hours.



Another conclusion is that reporting learning outcomes in terms of the CEFR levels (which is a form of criterion-referencing) makes it possible to report progress over time and to compare levels attained in different courses/languages much better than is possible in the still dominant norm-referenced grading practice in Finland.

A third conclusion is that the benefits of using the CEFR do not come cheap. A lot of effort has to be devoted to planning, execution, data analysis and interpretation. After an intensive period of development work there are now several tools that can be consulted and used, but even so the need for competence building and learning-by-doing should not be underestimated. On the other hand, reliance on external “experts” should not be overstated. Competence can be built up by making a commitment to a relatively long period of development work. A lot can be learned by studying how the CEFR has been used in other contexts.

Finally, I wish to express my personal perception of how the CEFR is viewed in Finland. It has been seen as a valuable tool in all national development of language education and also been found useful in international contacts and cooperation. It is seen as a reference tool, descriptive rather than prescriptive, both inviting and requiring thoughtful application by the users. While it is comprehensive it does not cover everything. Also, while it is the most useful tool developed in the recent past, it needs to be elaborated through international cooperation. In sum, both the CEFR and the ELP are good examples of international cooperation undertaken voluntarily and serving enlightened national self-interests. Contrary to some voiced criticism, it is not seen in Finland as an agenda for trying to enforce consensus or to exercise power. All of my forty years of involvement in the various CoE language project activities suggest that the ethos of the activities is built on sharing, consultation and cooperation.

References

- Alderson, J.C. (2005) Diagnosing Foreign Language Proficiency. The Interface between Learning and Assessment. London: Continuum
- Coste, D. (2006) Le Cadre européen de référence pour les langues: traditions, traductions, translations. Retour subjectif sur un parcours. In E. Piccardo (Ed.) La richesse de la diversité: recherches et réflexions dans l'Europe des langues et des cultures. Synergies No 1, 40-55
- Hildén, R. & Takala, S. (2007) Relating descriptors of the Finnish school scale to the CEFR overall scales for communicative activities. In A. Koskensalo, J. Smeds, P. Kaikkonen & V. Kohonen (Eds.) Foreign Languages and Multicultural Perspectives in the European Context. Berlin: LIT Verlag, 291-300. (pdf available on request at: sjtakala@hotmail.com)
- Kaftandjieva, F. , Verhelst, N. & Takala, S. (1999) A Manual for Standard Setting Procedure – DIALANG. Unpublished manuscript. (pdf available on request at: sjtakala@hotmail.com)
- Kaftandjieva, F. (2001) Standard setting in the Polytechnic assessment (AMKKIA) project of language proficiency. Unpublished manuscript. (pdf available on request at: sjtakala@hotmail.com)

- Kaftandjjeva, F. & Takala, S. (2002) Council of Europe Scales of Language Proficiency: A Validation Study. In: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies: Strasbourg: Council of Europe, 106-129
- Kaftandjjeva, F. & Takala, S. (2003) Development and Validation of Scales of Language Proficiency. In W. Vagle (Ed.) Vurdering av språgferdighet. Trondheim: NTNU, 31-40 (pdf available on request at: sjtakala@hotmail.com)
- North, B. (1995) The development of a Common Framework scale of language proficiency. Based on a theory of measurement. Thames Valley University, November 1995. Unpublished manuscript.
- Takala, S. & Katftandjjeva, F. (2002) Relating the Finnish Matriculation Examination English Test Results to the CEF Scales. Helsinki Seminar, June 31-July 2, 2006. Unpublished manuscript (pdf available on request at: sjtakala@hotmail.com)
- Takala, S. & Kaftandjjeva (2004) Using the Common European Framework: Some Finnish Experiences. In K. Mäkinen, P. Kaikkonen & V. Kohonen (Eds.) Future Perspectives in Foreign Language Education. University of Oulu, Faculty of Education Research Reports 10, 45-54. (pdf available on request at: sjtakala@hotmail.com)
- Takala, S. (2006) National syllabuses for foreign languages in Finland: Tradition and reform. In M. Berndtsen, M. Björklund, C. Fant & L. Forsman (Eds.) Språk, lärande och utbildning i språk i sikte. Festschrift tillägnad professor Kaj Sjöholm. Pedagogiska fakulteten vid Åbo Akademi, No 20, 217-225. (pdf available on request at: sjtakala@hotmail.com)
- Toropainen, O. (2010) Assessment of attainment in Finnish in the basic school. Helsinki: National Board of Education. (in Swedish)
- Tuokko, E. (2007) What level do pupils at the end of the basic school reach in English? Jyväskylä: University of Jyväskylä (in Finnish)
- Tuokko, E. (2009) What is the level of Swedish at the end of the basic school? Helsinki: National Board of Education. (in Finnish)

sjtakala@hotmail.com

Sauli Takala obtained his PhD in 1984 at the University of Illinois at Urbana-Champaign focusing on vocabulary learning in EFL. He took an active part over 40 years in the national research and development work on language education in Finland and participated in Nordic cooperation in this area, which was very lively in the 1970s and 1980s. He coordinated the IEA International Study of Writing in the 1980s and helped to plan and coordinate the EU DIALANG project in the late 1990s. He has had a long association with the Council of Europe modern language project and is currently consultant for its European Centre for Modern Languages. He is a founding member of EALTA and served as its second president.

Improving Classroom Assessment by Using the CEFR

Sanja Wagner

Erich Kästner-Schule, Darmstadt/Germany



Introduction

More and more educational systems move in the direction of increased accountability and place greater assessment demands on teachers. They not only give grades for pupils' performance in tests, they have to organize the learning process so that at the end of secondary school the pupils meet the national standards, which are related to the CEFR in all European countries. There is a need for a shift from the "assessment of learning" to "assessment for learning", an ongoing process of evaluation of the learning outcomes by pupils themselves as well as by teachers in order to promote future learning and language growth.

Assessment for Learning

Sharing the job of assessing the learning outcomes, as well as the learning process with the pupils, makes a metacognitive discourse about learning necessary, it eventually becomes central to classroom practice. Checklists from the CEFR and "rubrics" provide a common language as well as valid descriptors for different language skills. They promote understanding of goals and criteria and thus help learners to know how to improve. Assessment for learning in this way focuses on how pupils learn and not only on what they have learnt.

Having heard about the CEFR in 1998 I started to change my classroom practice in order to teach according to the CEFR's key ideas to make the teaching and learning process transparent and to facilitate autonomous learning through self-monitoring and self-assessment:

(...) it is one of the principal functions of the Framework to encourage and enable all the different partners to the language teaching and learning processes to inform others as transparently as possible not only of their aims and objectives but also of the methods they use and the results they actually achieve. (CEFR 2.3.1.)

I was encouraged to continue my experimental work on the new assessment tools in English classes when the national standards became the issue to discuss and implement in German schools in 2004. Beside standards, competence has become the buzz word in educational debates. As language teachers we are the lucky ones, having a huge bank of valid descriptors,

since there is an international agreement that CEFR is the reference for all educational boards in Europe. Examples from my teaching practice discussed in this paper show the potential of criterion-orientated checklists and grids for the assessment of task-based learning in young learners' classroom at levels A2 to B1/B2.

Work Plan

The first step was to introduce a “work plan” which includes all the learning activities during the following two to three weeks. Thus the pupils know what they are expected to do and how, which makes the teaching and learning process transparent. The plan includes various ways of and tasks for learning vocabulary, for reading and listening comprehension, for writing and communication as well as for small projects. There is also a section focusing on language awareness.

Working to a plan of this nature the teacher is able to stimulate reflection on

- what has been done so far,
- what the students have learnt,
- where the problems can be identified
- what students like or dislike about English lessons.

Topic: Extreme sports		Portfolio-self-assessment		→ Listening comprehension → Reading skills → Present perfect → Prepare a talk on extreme sports	
Class: 9 E-NHG-5B		m.88888			
Nr	Subtasks and tasks	How?	done?	checked	
1-LC	Thrills: Listen to what people say about the five thrilling activities. Take notes! Then tell your partner about these activities in the order you have heard them. You may look at the pictures in the book.	TB 28a	class alone	o	o
2-RC	Read the text on Extreme sports and soft adventure. Tell your partners in your group what the text is about (summary of the text).	WS	alone group	o	o
3-RC	Been there, done that! Listen to the CD and reading. When the CD stops make some notes on what is important to know. Don't write complete sentences!	TB 31a	class	o	o
4a	Check your understanding and do ex. 5b! Don't look back to the text. Can you correct the wrong ones? When we check the results, write down how many right answers did you have!	TB 32/35a	Alone and be correct tested	o	o
5a	Now go back to the text and correct all your answers in the end and check your solutions with the text. Headline the text again and then complete the task in WB 15. Don't copy from the text. Use your memory!	WB 15/31 and 52a	Alone check with partner	o	o
6-Write	Now you know all about the bungee jump story! Write a short news about it in about 10 sentences.		alone	o	o
7-Voca	In this text there are many words you must learn and you can use a lot. This time you collect all the verbs in one list all the nouns in another one all the adjectives in the third list The lists must be complete!	TB 18/31 to 37a	alone partner	o	o
8-Voca	How the words are made! Use the syllables to find the words!	WB 19/35a	o	o	o
9-Activity	Great photos! How adventurous are you? Answer the questions truly and count the points. Talk about your result in your group. Report to the class.		alone group class	o	o
10-LJFa	What have you done? When do we use present perfect? Look up the rules in WBW and do the exercises in WB.	WB 17/34	alone	o	o
11-LJFa	Since and for... what's the difference?	19/34, 35a	alone	o	o
12-LJFa	Write an instruction how and when to use present perfect on the orange card. Put it into your grammar file!	WBW	partner	o	o
13-Talks	Prepare a talk about an extreme sport of your choice! Search the Internet or magazines if you need more information!		alone partner class	o	o
14a	Selbstbeurteilung: Wie gut bist du in Schreiben? Lesen und Hören? Portfolio Stufe B1		alone	o	o

For students the “work plan” is a scaffold which supports their engagement in learning and in negotiating while the teacher is able to stimulate reflection on learning progress, achievements, problems as well as likes and dislikes when learning a language. Pupils learn how to learn a new language in general as well as developing their capacity to plan, monitor and evaluate their own learning and build up a dossier of their best work.

Observing the students while they are working on different tasks, the teacher gets to know more about the students and their ways of learning a language. Thus he is able to respect individual learning styles as well as individual needs, aptitudes and interests without giving up his goals set out in the syllabus. For the gifted and talented the unit plan offers the possibility to go far beyond the work in the classroom. Before the teacher starts to develop a “work plan” on a specific topic it is important to reflect on the following aspects:

- Which learning techniques and social forms do I want to introduce/ reinforce?
- How much room do I give to my pupils for autonomous learning?
- What do I want to have evaluated by a self-assessment grid?
- When and how often do I do this?
- What examples are there to facilitate my job?

A workplan is a scaffold for introducing and fostering experiential learning, self organized/ autonomous learning, monitoring and self-evaluation. The research on language acquisition tells us that there is no such thing as step-by-step learning following the so-called grammar progression (which is at the heart of most course books in Germany). On the contrary – learners must be exposed to as much target language as possible like children learning their mother tongue. This “bath of language” is equally essential for the acquisition of language for second-language learners. Michael Swan explains the importance of the extensive language input as follows: *“They need to be exposed to quantities of spoken and written language, authentic or not too tidied up, for their unconscious acquisition processes to work on.”* (Swan 2006) Within a workplan pupils are exposed to a great choice of language materials and they can choose what to do and how to do it individually.

Accuracy against Communicative Effectiveness

The language acquisition research project by W. Bleyhl states an “ultimate paradox” in language teaching, saying that the more we focus on formal aspects and accuracy and the more we simplify and break up the language into “eatable pieces” the less the learners acquire language competence. (Bleyhl pdf) In other words the PPP method is totally contra-productive, as well as teaching only vocabulary (e.g. English=German) or grammar, even if however this is very easy to teach and to assess. Above all, teachers are praised for testing vocabulary and grammar on a regular basis, which even parents consider most important in language classes.

Especially pupils at the lower intermediate level, who start processing language on their own, are very likely to make mistakes which are typical for the stage of their learning process. Teachers often judge the performance by focusing on accuracy, on mistakes and consequently, even if the pupil tries very hard to use the target language and successfully communicates his thoughts, he gets penalized for mistakes. This is a very frustrating experience and a lot of pupils draw back into silence instead of trying and experimenting with the new language. In English classes we must create a safe learning environment, where

these pupils can experiment with language and use their interlanguage, which of course is not always correct and fluent but is a necessary stage in the process of language acquisition. Being familiar with the CEFR linguistic scales at level A2 to B1, teachers get much more tolerant and focus more on communicative effectiveness and the receptive skills like reading and listening, where pupils usually achieve the target level with ease.

Task-based Language Teaching and Learning

There is a need to change classroom practice so that the process of second language acquisition is embedded in subject-area learning, enabling pupils to build up topic knowledge and carry out different activities for which they develop useful strategies and skills (CEFR p.137). Task-based learning and project work enable pupils to learn the language in this way. They work in a group on a content-based task, present the outcomes and in doing this use the target language. This is a very complex performance. Assessing pupils' performance on a task means not only evaluating the final product (e.g. a reader, poster, multimedia presentation...), but also monitoring and appraising a very complex process over a period of time. Many teachers avoid this, just because they are not aware of the tools available, such as checklists and rubrics for the possible areas of assessment. These are:

- expanding their knowledge on a specific topic
- working collaboratively in a team
- presenting the results and using visual aids
- using and improving their language skills
- developing learning and thinking skills

Working on a project in a team focuses on a topic (Media, Geography, History, Science...) where the target language is the means of communication, oral and written. Pupils are doing things in the target language all the time, be it by reception or production.

- Pupils are **exposed to extensive language input**, searching for the information in authentic sources (books, magazines, Internet, films, lyrics)
- They produce **extensive output**, writing notes, summaries, drafts
- Editing their material and rehearsing involves **intensive output**
- There is plenty of opportunity to negotiate language use: "how to put it right", what is the most appropriate way of expressing one's thoughts

- There is a need to talk about sentence structure or tenses, thereby dealing with grammar in a meaningful way (**analysed input and output**)
- There is the opportunity to work collaboratively, helping each other and learning from each other, to show strengths and admit weaknesses
- Working on a project is a challenging, meaningful and authentic experience in the target language

It offers plenty of opportunities for self-organized learning and (self-) monitoring / assessment and the final product – in most cases a presentation – is an invaluable but rare situation to assess oral skills in the classroom.

Assessment with checklists and rubrics

For the teacher there is the crucial question: How to assess pupils' performance when language is not the only focus of assessment, because they are working on a task or even on a project about history, geography or other cultural or political issues?

In order to make the monitoring and assessment easy to handle in classroom setting and transparent for the learners it is important to design checklists and link them to the workplan so that students can join in this assessment process themselves, monitor and assess their own progress and thus start to organize their learning process themselves. The teacher uses the same evaluation tools for his assessment.

Topic: []		Die Europäische Union / die Sprache		
What I can do		no problem	with some mistakes	cannot work on this
listening				
speaking				
reading				
writing				
Phonetics	I can use the new words and phrases and pronounce them clearly but with a slight accent.			
Spelling	I can write the new words and phrases mostly without any mistakes.			
Grammar	I can make simple sentences correctly, the word order and the tenses are (mostly) correct.			

Topic: []		Die Europäische Union / die Sprache			I can do it!	I must practice this!	It is too difficult for me!
Language-skills	Activities, tasks, projects	very good	good	practise	practise	difficult	
listening							
speaking							
Spoken interaction							
mediation							
Reading comprehension							
Spelling	I can write the new words and phrases mostly without any mistakes.						
Grammar							
Vocabulary							
Pronunciation	I can use the new words and phrases and pronounce them clearly and neatly without any accent.						

The checklists mostly consist of positive “can do” statements taken from the CEFR which reaffirm the achievements of the individual student and strengthen his self-esteem.

Report on the reading logbook on ->

Title:

Student Name:

CATEGORY	4 excellent	3 good	2 satisfactory	1 needs-improvement	Points
Grammar and vocabulary	There are nearly no mistakes and the language is ambitious.	There are some mistakes and the language is often ambitious.	There are more mistakes but non-impeding and the language is mostly very simple.	There are many mistakes the language used is very simple and often difficult to understand.	
Neatness and Effort	The logbook has no distracting errors, corrections or erasures and is easily read. It appears the student spent a lot of effort getting things just right. It looks like the author took great pride in it.	The logbook has almost no distracting errors, corrections or erasures and is easily read. It appears the student worked hard on it. It looks like the author took some pride in it.	The logbook is fairly readable but the quality is not too good on some parts. Student devotes some time and effort to the writing process but was not very thorough. Does enough to get by.	Very messy and hard to read. Student devotes little time and effort to the writing process. Doesn't seem to care. Or the student did hardly work on it.	
Writing-Process	Student devotes a lot of time and effort to the writing process (prewriting, drafting, reviewing, and editing). Works hard to make the logbook wonderful.	Student devotes sufficient time and effort to the writing process (prewriting, drafting, reviewing, and editing). Works and gets the job done.	Student devotes some time and effort to the writing process but was not very thorough. Does enough to get by.	Student devotes little time and effort to the writing process. Doesn't seem to care.	
Creativity	The logbook contains many creative details and/or descriptions that contribute to the reader's enjoyment. The author has really used his imagination.	The logbook contains a few creative details and/or descriptions that contribute to the reader's enjoyment. The author has occasionally used his imagination.	The logbook contains a few creative details and/or descriptions, but they distract from the story. The author has tried to use his imagination.	There is little evidence of creativity in the logbook. The author does not seem to have used much imagination.	
Requirements	All of the written requirements (of tasks) were met.	Almost all (about 80%) of the written requirements were met.	Most (about 65%) of the written requirements were met, but several were not.	Many requirements were not met.	
Date	25. May - 2008				

But there is often a need for a more detailed description of students' performance, based on explicit criteria, which is provided by using “rubrics”. For teamwork, reading logs and presentations with the help of visual aids I use rubrics, mostly adapted from the free online tool available on the internet “<http://rubistar.4teachers.org>”. The advantage of rubrics is that they give explicit description of a performance usually on four different levels – so the pupils learn the criteria for a very good performance and realize why theirs is or is not that good. On the one hand rubrics provide transparent criteria of evaluation, on the other hand they help the pupils to set themselves new goals because they now know

what to improve.

I was able to work with the same group of learners for 6 years, thus developing autonomous learning step by step, observing the pupils as well as discussing with them the pros and cons of the new approach. For the last three years, we used the *European Language Portfolio* published by Diesterweg. There was no need to correct any self-assessment in the language passport, however it was a long and difficult road in a school system and in a society where self-assessment, ownership and self-directed learning are just starting to emerge. However introducing this new teaching and learning practice in a class, which was not used to it, is very hard work, takes a long time and may even fail. For pupils it is much easier to learn new words and grammar rules by heart and to fill in gaps in exercises than to engage in tasks, use and produce language and above all reflect on their language growth. They must learn to feel free to take risks and consequently we must encourage them time and again to go on talking without fear of making mistakes and to continue learning by giving them meaningful tasks which they can manage and by giving them positive feedback using valid and transparent criteria.

The “can do” statements, “rubrics” and portfolio assessment bring a new culture of assessment

and evaluation into the foreign language classrooms, which again contributes to life-long language learning and offers “opportunities to acquire independence and autonomy as learners (....) and encourages co-operation and other social values.”(Hayworth 2004)

References

- Bleyhl, W: Die Defizite des traditionellen Fremdsprachenunterrichts oder Weshalb - endlich - ein Paradigmenwechsel, eine Umkehr, im Fremdsprachenunterricht erfolgen muss (pdf from <http://creativdialogues.lernnetz.de/docs/1.4Expertenstimmen.pdf>)
- Hayworth, F. (2004) *Why the CEF is important*. In K. Morrow: Insights from the Common European Framework. Oxford: OUP. 13
- Swan, M.: *Two out of three ain't enough – the essential ingredients of a language course* (2006) In Harrogate Conference Selections IATEFL
- Wagner, S.: (2006) *Working with the CEF*. In SIG Independence Newsletter, Issue 39
- Wagner, S.: (2006) *Towards autonomous learning*. In Harrogate Conference Selections IATEFL
- Wagner, S.: (2008) *Kompetenzorientiert bewerten im Englischunterricht*. In *Praxis Schule 5-10*, Heft 5/Okttober2008, Westermann
- Material to download: <http://quiclr03.teamlearnlive.de/b-1-egt> (read EQT)

sanja-wagner@web.de

I teach mixed-ability classes at the state comprehensive school in Darmstadt near Frankfurt. As a teacher trainer my main focus is on the implementation of CEFR and ELP in English classes and on improvement of literacy skills. I have been developing portfolio assessment sheets as well as test formats for the new course books *Notting Hill Gate*, published by Diesterweg.



Judith Mader
IATEFL TEA SIG
Newsletter Editor



Zeynep Urkun
IATEFL TEA SIG
Events Coordinator

'Putting the CEFR to Good Use' is a collection of selected articles written by the presenters of the IATEFL TEA SIG and EALTA conference in Barcelona, Spain held in October 2010.

The impact of the Common European Framework of Reference for Languages (2001) and of the Manual for Relating Examinations to the CEFR (2009) in the field of language testing and assessment has resulted in a growing number of research programs, linking projects and training endeavours. However, different constituencies in different contexts of use with different resources create different scenarios which require tailor-made approaches in terms of improving current practice, and managing change, competence building in the area of using the CEFR not only in exam contexts but also in classroom assessment, increasing the quality of test development and test administration procedures, developing procedures that guarantee transparency and accountability, encouraging the development of both formal and informal national and international networks

The aim of this conference was to look at how professionals in the field have addressed these issues, and to exchange ideas on how different constituencies can cooperate in order to improve testing and assessment practice(s) in Europe.

June 2011

IATEFL TEA SIG Committee and EALTA are proud to present this unique publication to the International ELT community as an e-book.

The IATEFL Testing, Evaluation and Assessment Special Interest Group (TEA SIG) is one of the fourteen SIGs at IATEFL. It was established in 1986, with the aim of reaching those who are interested in the process and product of learning English as a second and foreign language, in testing, evaluation and assessment in ELT.

EALTA (European Association for Language Testing and Assessment) is a professional association for language testers in Europe. The purpose of EALTA is to promote the understanding of theoretical principles of language testing and assessment, and the improvement and sharing of testing and assessment practices throughout Europe.

You can find out more about the IATEFL TEA SIG at: <http://tea.iatefl.org/>

You can find out more about the EALTA at: <http://www.ealta.eu.org/>