CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 63

March 2016

ISSN 1756-509X

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 63 / March 2016
A quarterly publication reporting on learning, teaching and assessment

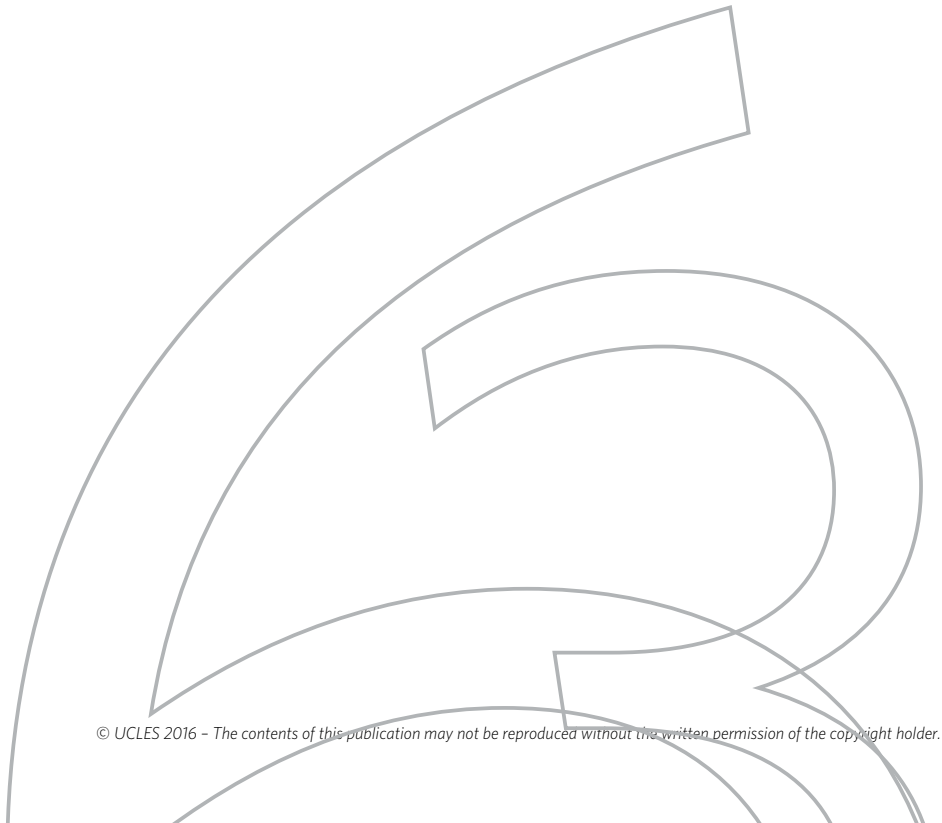
Printed in the United Kingdom by Canon Business Services

Research Notes

Issue 63

March 2016

Contents



Editorial

Welcome to issue 63 of *Research Notes*, a quarterly publication offering updates on the Cambridge English approach to learning, teaching and assessment, and its impact on key stakeholders. The focus in this issue is collaboration with European projects.

The opening article by Saville and Gutierrez Eugenio distils the findings of the European Commission's 2015 *Study on comparability of language testing in Europe*. As a followup to the Council Conclusions of May 2014 to promote multilingualism, this report investigated to what extent EU Member States' tests of language competence were comparable and how such parity could be achieved. The project was conducted to a tight timeframe and based largely on comparisons of existing data compiled by Eurydice on Member States' foreign language testing systems. Data on test materials and test performance was collected from IEG members and although this proved to be limited, reliability of results was not affected. A content analysis tool was developed for comprehension of the construction and validity of the language tests, partially guided by the Council of Europe's Common European Framework of Reference for Languages (CEFR). Analysis of each existing test in each country and in each International Standard Classification of Education (ISCED) level was conducted by expert content analysts. The results showed that the tests were not easily comparable in terms of constructs, inferences and measurements characteristics, which seriously decreased any chances of reaching test results which could be comparable across countries. These findings did however inspire a fresh set of measures which could be implemented at a national level to meet the demands of increased social mobility and globalisation, including improved understanding of the CEFR and greater focus on enhancing students' language learning skills.

One of the most important innovations arising from the above project was the 'No More Marking' website, a psychometric tool designed to overcome the problems of test comparability, described by Jones in the next article. The notion of comparative judgement (CJ), where judgements are paired in order to form an accurate ranking of test performance, strongly informed the design of the website. Jones presents findings which attempted to validate the level of difficulty in ISCED 2 and 3 tasks of French reading and writing and English reading and writing. He concludes that this tool may obviate the need for pretesting while maintaining standards across a variety of tests and raters.

North and Panthier continue the theme of refining descriptors to guide language testing educators throughout Europe. Starting with an introduction that corrects some the popular misconceptions regarding the CEFR, the authors then guide us through some of the arguable forerunners of the publication, such as the Threshold level and the Cambridge English suite of publications it inspired, and the CEFR's first appearance in 2001. Updates to the CEFR descriptors are set in the context of recent initiatives for plurilingual and intercultural democratic citizenship, which view the learner as an individual with a repertoire of languages which permeates their educational and professional development. The project to update descriptors (which is near completion at the time of press) is delineated over two phases: the first to refine 2001 descriptors, the second to develop descriptors from scratch. The latter are presented here as mediation activities and strategies. The following article, by North and Docherty, elaborates on the second phase of the project briefly described by North and Panthier, explaining how fresh descriptors were created for mediation, a communicative language activity that incorporates elements of the other major communicative activities reception, interaction and production. The development and validation processes are described in full, and the authors conclude that refinements in validating plurilingual competence will be required and highly welcomed.

The final two articles demonstrate how project findings are shared and implemented in diverse contexts. Martyniuk begins with a report on the 2015 'Learning Through Languages' conference of the European Centre for Modern Languages (ECML). Plurilingualism is at the heart of the ECML mission. Martyniuk summarises the projects that illustrate this intercultural and interdisciplinary notion, as well those that offer practical advice to educators. His conclusion guides us to how such current initiatives will develop over 2016–2019. In the final article, Fiona Barker reports on the sixteenth English Profile Network Seminar held in Cambridge in February this year, which celebrated some of the highlights of English Profile over the last decade.

The European Commission's 'Study on comparability of language testing in Europe' (2015)

NICK SAVILLE RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT
ESTHER GUTIERREZ EUGENIO RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Setting the scene: The EU language policy context

Europe is known for its rich diversity at all possible levels, which the European institutions strive to maintain while encouraging mutual understanding among its citizens. Languages are essential to achieve this key EU aim of 'united in diversity' since they help build bridges between cultures and societies and enable European citizens to move freely throughout Europe, be it for personal, academic, or professional reasons. Languages are also crucial to ensure participation in society by all European citizens, including those from marginalised language groups. More importantly, languages have become a fundamental basic skill to boost employability, particularly among young people, which in turns leads to economic growth and better living standards.

The European Union does therefore have a number of significant reasons to promote language learning and linguistic diversity among EU Member States (EU MSs). However, it is important to remember that MSs are responsible for their own education policies, and the European Commission's role is solely to co-ordinate and support countries' efforts in this field, in line with the principle of subsidiarity, and to provide opportunities for MSs to share objectives and best practices with each other.

In the field of language education, most of the work that the European Commission has been undertaking over the past 15 years has been rooted in what is known as the 'Barcelona objective': the aim that all European citizens should be able to communicate in two languages other than their mother tongue. This goal was agreed at the Barcelona Conclusions of the European Council in March 2002 by all the EU's Heads of State and Government, and called for 'action to improve the mastery of basic skills, in particular by teaching at least two foreign languages from a very early age', and also for 'the establishment of a linguistic competence indicator in 2003'. The Barcelona Conclusions were motivated by the lack of reliable data to monitor progress in the development of language skills of EU citizens, and they have driven most efforts at national and European level to improve language learning and teaching across MSs since 2002.

As a result of the Barcelona Conclusions, and after detailed discussions with official representatives from MSs, the European Commission presented a plan to set up the European Indicator of Language Competences in 2005. This indicator would be used to monitor progress in language skills across EU MSs, which would be measured by administering standardised language tests to a representative sample of students aged 15 from across MSs. After an open tendering process, the consortium SurveyLang (led by Cambridge English Language Assessment) was commissioned in 2008 to deliver the first European Survey on Language Competences (ESLC), the data of which would then be used for the creation of this European Indicator of Language Competences. The 3-year project took place across 14 EU MSs, and showed both the need to improve secondary school students' language competences in order to achieve the Barcelona goal, and that there is a wide diversity of teaching and assessment methods being used across EU Member States (European Commission 2012a).

In light of these results, in their *Rethinking Education* communication (2012b), the Commission presented a proposal to create a European benchmark on languages which would enhance multilingualism across Europe and help monitor MSs' progress. This benchmark consisted of two indicators with the following two goals to be met by 2020:

- at least 50% of 15-year-olds attain the level of independent user or above of a first foreign language
- at least 75% of pupils in lower secondary education study at least two foreign languages besides their main language of instruction.

The Commission proposed monitoring the first indicator by regularly repeating the ESLC, and combining these results with national data from those Member States who did not wish to take part in this survey. However, the 'Conclusions on Multilingualism and the Development of Language Competences', adopted by the Council of the European Union in May 2014, rejected the Commission's proposal for the European benchmark on languages. Instead, the Council invited the European Commission to explore alternative measures to improve the quality of language learning, teaching and assessment within each MS as a way of actively promoting multilingualism. Additionally, the Council suggested examining the comparability of results collected through the different national assessment systems, and the potential for these results to help regularly monitor MSs' progress towards the Barcelona goal.

Aims: What the study intended to achieve

Following the recommendation put forward in the May 2014 Council Conclusions described above, the European Commission published a call for tenders to explore the feasibility of making use of existing national language

tests to assess language competences across all EU MSs. This study was conducted in the frame of this tender, and the aim was to critically assess the comparability of existing national tests of pupils' language competences at levels International Standard Classification of Education (ISCED) 2 (lower secondary education) and ISCED 3 (higher secondary education) in the 28 EU MSs. The aim of this project is therefore to critically assess the comparability of existing national tests of students' language competences in Europe at ISCED 2 and ISCED 3 levels. In order to fulfil this aim, there are two questions that need to be answered:

1. Are tests of language competences at the same ISCED level across all European countries testing the same thing? In other words, do all European countries have a common understanding of what language competence means and how it can be measured?
2. Even if the answer to the first question was that they actually do test the same thing, are they doing it reliably enough to claim that results could be comparable?

With these two questions in mind, it was possible to design the analytical framework that was used to compare the language exams in this project and which is described later in this article. This article also explains how this framework was operationalised in the Content Analysis Tool, which experts used to examine existing language tests and provided the most important part of the data leading to the final results.

Circumstances: Factors which limited the scope and implementation of the study

There were a number of circumstances which were set out by the Tender Specifications and which determined the scope of the study and the way in which it was carried out. These circumstances are considered in some detail in this section as they had an important impact on the planning and implementation of the project.

Structure of the study

The Tender Specifications stated the five main tasks that needed to be covered in the final report, and the methodology was designed and implemented with the goal of fulfilling the following five tasks (European Commission 2015:20):

Task 1: Assessment of comparability of the existing national language tests administered to secondary school students.

- Produce a critical yet constructive overview of comparability between different existing national or regional methods to assess language competences in Europe's secondary schools.

Task 2: Proposals for ex-post adjustment that can increase the comparability of existing results of language tests at national level.

- Identify and describe in detail proposals for measures and methodological procedures potentially needed to adjust for methodological differences in the results of existing national tests, in order to present the country aggregates in a coherent and meaningful European overview. This task directly concerns those jurisdictions that already have a national or regional system of language testing.

Task 3: Proposals for development work that could be undertaken at the national level to increase comparability of existing language tests.

- Identify and describe in detail proposals for development work that could be implemented by EU MSs already having national or regional language tests, in order to increase the European comparability of their data. This task directly concerns those jurisdictions that already have a national or regional system of language testing.

Task 4: Proposals for development work that could be undertaken at national level by MSs not having a system for language testing and interested in developing one.

- Identify and describe in detail proposals for development work that could be implemented by EU MSs not having implemented national or regional language tests yet, with an approach that yields results comparable to other European jurisdictions.

Task 5: Comparative overview of existing country data on language testing.

- Compile an overview of country data on language testing.

Timeframe

The study was initially intended to last five months, from the signature of the contract to the delivery of the final report. This timeframe included several meetings between the Cambridge English team and the project team at the European Commission, as well as several meetings with the Indicator Expert Group on Multilingualism (IEG), which comprised representatives from the Ministries of Education of the different EU MSs. This timeframe also included a 3-week period for IEG members to provide feedback on the draft final report, which fell in August. For practical reasons, the European Commission decided to extend this period until mid-September, which also meant extending the length of the project by one additional month. However, this did not impact on the development of the project as the draft final report still had to be submitted at the end of July on the same date as initially agreed. The European Commission's main objective in setting this timeframe was to be able to present the results at the European Day of Languages event in Brussels, which took place on 25 September 2015, and at which Cambridge English was invited to provide a short presentation about the project.

The timeframe with the main milestones is provided in Figure 1. As it can be observed, the majority of the work had to be conducted in the 6-week period between the

approval of the Inception Report by the Commission in mid-May and the end of June, when the report had to start being drafted in order to meet the submission deadline on 24 July. In these weeks, all the data had to be collected, organised, coded, analysed and the results written in the form of the Draft Final Report. This tight timeframe to undertake all the tasks leading to the final report was a key limiting factor which had to be carefully considered when determining the practical scope of the study and when designing the methodology.

One of the main risks was underestimating the time needed for the workload and thus delaying final delivery of the project. Involving more people in the project did help with the workload, especially in regard to data management, but the short timeframe did also limit any training or mentoring opportunities. This means that only very capable staff with a thorough knowledge in the field and familiarity with the procedures could be involved in the project as there was no time to train and support any new project team members once the project had started. Therefore the agreed priority throughout the project was to deliver on time and adapt the scope to cover as much as realistically possible within the given timeframe and with the best human resources available. This circumstance had an important impact on other parts of the project, such as the nature and amount of data collected or the types of analyses conducted and presented in the final report.

Geographical coverage

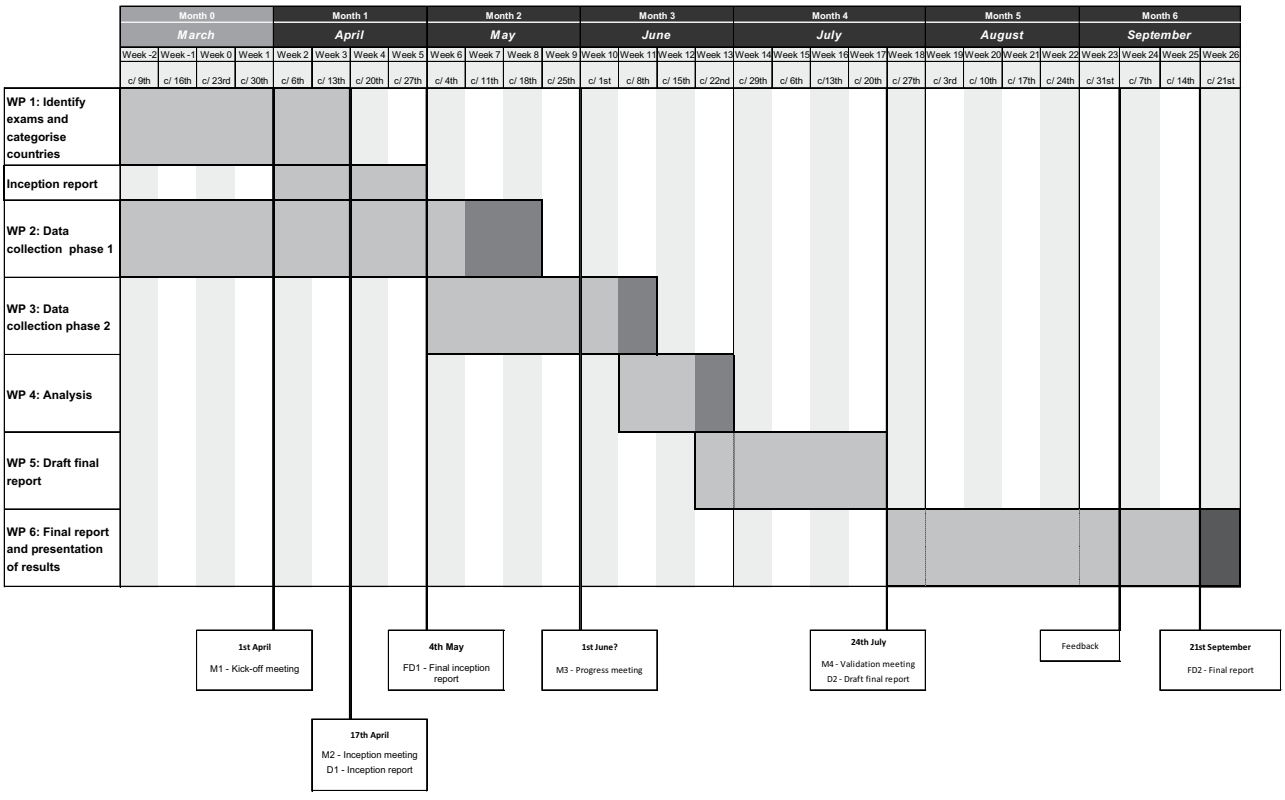
The Tender Specifications required the study to cover all 28 EU MSs. However, some EU MSs such as the United Kingdom and Belgium have independent educational authorities for each part/region with different language testing systems. For the purpose of this study, these parts/regions were considered as separate entities, which led to looking at existing language exams in the 33 educational jurisdictions identified across the 28 EU MSs.

Nature and amount of data collected

The Tender Specifications stated very clearly the nature of the data that had to be collected and analysed in this study. The goal was to use existing data from national language tests, which is defined as exams 'generally organised at the national or regional level, implemented in secondary schools and funded by national or regional budgets for education and training' (European Commission 2014:14). These exams are usually organised by central/top-level public authorities, although in some cases they are designed and run at a regional level, as explained in the previous section.

According also to the Tender Specifications, the languages included in this study are 'languages that are not the main language of instruction; or, in other words, [competences] in one or more foreign languages' (European Commission 2014:14). Considering the current focus on increasing mobility and access to jobs in other EU countries, only EU official languages that are used in at least one other EU MS

Figure 1: Timeline of the project*



*M = Milestone; D = Deliverable; FD = Final deliverable

were included in this study. For each jurisdiction, only those languages studied by more than 10% of secondary education students (according to Eurostat 2013, 2014) were considered. Table 1 shows the languages selected for each jurisdiction.

Table 1: Foreign languages most taught in each jurisdiction and included in this study

Jurisdiction	First foreign language	Second foreign language	Third foreign language	Fourth foreign language
Austria	English	French		
Belgium FR*	English	Dutch	German	
Belgium GE*	French			
Belgium NL*	French			
Bulgaria	English	German		
Croatia	English	German	Italian	
Cyprus	English	French	Italian	
Czech Rep	English	German		
Denmark	English	German		
Estonia	English	German		
Finland	English	Swedish	German	
France	English	Spanish		
Germany	English	French		
Greece	English	French	German	
Hungary	English	German		
Ireland	French	German	Spanish	
Italy	English	French	Spanish	
Latvia	English	German		
Lithuania	English	German		
Luxembourg	German	French	English	
Malta	English	Italian	French	
Netherlands	English	German	French	
Poland	English	German		
Portugal	English	French	Spanish	
Romania	English	French		
Slovakia	English	German		
Slovenia	English	German		
Spain	English	French		
Sweden	English	Spanish	German	French
UK England	French	German	Spanish	
UK Northern Ireland	French	German		
UK Scotland	French	German		
UK Wales	French	German		

*FR = French-speaking; GE = German-speaking; NL = Dutch-speaking

This project did therefore not collect any raw data about students' proficiency in foreign languages but rather tried to collect and critically assess the comparability of existing data collected through centralised language exams run at a national or regional level throughout the EU MSs. In order to explore the comparability of the data collected, this study also had to collect significant data about the data collection instruments used in each case, i.e. about the language tests used in each jurisdiction.

In order to facilitate this task, and in view of the tight timeframe given for the completion of this project, the European Commission facilitated access to the preliminary results of another project which was being conducted in parallel by Eurydice, a network of national experts from 37 European countries who provide information on national education systems to facilitate European co-operation and the production of cross-country studies based on reliable national data. Their work is co-ordinated by the European Commission, and at the time this project was conducted they were compiling an inventory of national language testing systems across Europe. Access to their preliminary findings was fundamental as it provided reliable and trusted information about the existing language testing systems in each jurisdiction, as well as some details about the design, administration and validation of the language tests. On the basis of the information provided by Eurydice, which was double checked with the members of the IEG on Multilingualism, 133 language examinations (33 jurisdictions, 28 EU MSs) were identified as relevant for this comparability study. Out of these 133 language examinations, 77 were at ISCED 2 level and 56 were at ISCED 3 level. Table 2 shows the number of exams testing each of the languages in this study, and Appendix 1 of the Final Report offers a detailed list of the national exams included in this study, as well as the reasons why certain exams had to be excluded.

Table 2: Total number of exams per ISCED level and number of exams testing each of the languages included in the study

Level	ISCED 2	ISCED 3
English	29	23
French	15	13
German	24	15
Spanish	6	3
Italian	2	1
Swedish	0	1
Dutch	1	0
Total	77	56

The data received from Eurydice did not include however samples of test materials, test performance or any other supporting documents which were needed to critically assess the comparability of results across these examinations. This data had therefore to be obtained by the project team, mainly through close liaison with the relevant IEG member or by asking in-country experts for advice on how and where to find the relevant documentation about the examinations.

The data collection was challenging as it had to be conducted during the busiest period of the year for national assessment boards, who were best positioned to provide the required information. In some cases, accessing the relevant documents would have taken weeks or simply have been impossible to obtain due to the confidentiality of the materials. It also proved difficult to gain access to any samples of performance (i.e. samples of students' performance in writing and speaking tests), and hardly any jurisdiction was able to provide this data, which would have been extremely helpful and allowed for a much more meaningful set of results. This lack of access to the relevant data in the pre-established timeframe was an important limitation of the study, and the analysis could only be conducted on the basis of the data obtained. However, this does not affect the reliability of the results, and important patterns and issues emerged which were common to all the examinations included in this study.

Collaboration with EU Member States

One of the European Commission's main concerns was to ensure the transparency and accuracy of the data collected and of the results obtained from its analysis. For this reason, the Tender Specifications set out very clearly that the project team would have to collaborate very closely with the members of the IEG. These members were all experts in language education and/or language assessment working for the Ministries of Education, National Statistical Offices or National Assessment Boards in their respective jurisdictions.

The European Commission organised two meetings with the IEG throughout the duration of the project: one at the very beginning to explain the methodology of the study and how their support and input would be essential to ensure the accuracy of the results. After the first meeting, the project team contacted each of the IEG members from the 33 jurisdictions and liaised with them to determine the exams that should be included in the study, the languages, and to discuss access to the additional test materials and supporting documentation (test papers, samples of students' performance, supporting documentation regarding the tests e.g. procedures for the creation and administration of exams, training materials for item writers and raters, national results, etc.). The second meeting took place at the end of the project after the IEG had provided written feedback on the draft final report to discuss with the European Commission and the project team regarding their general impressions and the impact the results could have in their different jurisdictions.

Implementation: From the analytical framework to the data collection and analysis

Analytical framework

In order to determine the comparability of the results from national language tests, it was first indispensable to look at the instruments used to collect these results, i.e. the actual language tests. The comparison of such a large number of language tests required a systematic approach to comparability through the development of an appropriate analytical framework. In order to do this, three major dimensions of comparability were identified:

1. The *construct* dimension relates to the *validity* of the assessment, i.e. how the language proficiency constructs are defined and tested. Within this dimension, the following aspects were considered:
   - the purposes which language education is to address
   - the way in which the different language skills are defined
   - how progress in acquiring those skills is measured.

   These aspects are important for comparability purposes because jurisdictions may differ in the way they conceive, prioritise and implement each of these aspects.
2. The *assessment* dimension relates to the *reliability* of the assessment, i.e. the technical features of item design, method of test delivery, frequency of assessment, etc. This dimension regards how the language proficiency constructs are operationalised and measured through test tasks and other forms of evaluation. This dimension is important for comparability because of two main reasons:
   - Test tasks may not be valid measures of the intended constructs, i.e. despite using the same definition of a given construct, the way this construct is measured may vary and this measurement may be more or less valid.
   - Tests' reliability may vary across different jurisdictions as well as across test sessions within the same jurisdiction. If test reliability is low, then the reliability of the data collected through these tests will be less meaningful and therefore less useful for comparability purposes.
3. The *performance* dimension concerns evidence: to objectively compare students' language proficiency across jurisdictions, data on students' performance would be required. This data can be in the form of:
   - responses to writing tasks and recordings of speaking tasks, annotated to show the marks awarded
   - tables of test scores from objectively and subjectively marked tests, showing the profile of achievement, and the interpretation of standards attributed to scores.

The CEFR as a framework for comparison

The Common European Framework of Reference (CEFR, Council of Europe 2001) provides a useful and systematic framework for comparison across different language learning, teaching and assessment systems. For this reason, the role and usefulness of the CEFR were highlighted throughout this study and provided an additional layer of comparability which was incorporated into the main analytical framework.

There are a number of reasons why the CEFR was considered as a key instrument for the comparison of national language tests. The first reason is that the CEFR currently acts as a familiar point of reference in language education in Europe, and has been referenced widely when defining the goals of language education, professional training of teachers, curriculum development, and as a scale for reporting learning outcomes. It provides a detailed discussion of how languages may be taught and learned as tools for communication, and presents a framework of levels, which sets out to enable a broad comparison of language-learning programmes and purposes. The CEFR also provides a relevant model which incorporates different perspectives on the nature of language learning while highlighting the importance of an action-oriented approach which understands the development of language skills as a result of motivated interaction within society. Finally, the CEFR can also be understood as a measurement construct where the descriptors for the six levels of proficiency (A1 to C2) were developed through an empirical study which used item response theory applied to tables of Can Do statements. These levels and descriptors have been adopted and developed by several examination bodies, primarily Cambridge English Language Assessment and the Association of Language Testers in Europe (ALTE), as a scale for reporting levels of achievement in language tests. The CEFR was also the reporting scale used in the ESLC and the CEFR levels were used as indication of language achievement in schools across Europe. For this reason, this

study has also anchored to the same CEFR scale used in the ESLC and strongly recommended the use of the CEFR levels to bring national language examinations into alignment.

Operationalising the analytical framework

In order to comprehensively assess the comparability of results from the different language tests across jurisdictions, the three above-mentioned dimensions would need to be explored. While elements regarding dimensions 1 and 2 (i.e. constructs and assessment) can be captured to a great extent through a careful analysis of the test materials and supporting documentation, dimension 3 (i.e. performance) requires accessing and examining data on students' performance, i.e. samples of writing and/or speaking, and test scores from the different exams with their respective interpretation. In the case of this study, and given the circumstances explained in section 3 above, the data necessary to address dimension 3 proved extremely difficult to gather, and Jones (this issue) presents in more detail the type of procedure which could be applied in case this data becomes available in the future.

Dimensions 1 and 2 (i.e. constructs and assessment) were operationalised according to Kolen and Brennan's (2004) 4-level comparability framework, which identifies:

1. **Inferences** (the intended purpose of a test within the broader education context).
2. **Constructs** (the understanding of what is being measured).
3. **Populations** (the characteristics of the students taking the exams).
4. **Measurement characteristics/conditions** (the characteristics of the test as a measurement instrument and the conditions in which it is administered).

This framework was further expanded and operationalised, in particular with reference to the ALTE Checklists, which were used at a lower level of detail to reveal the constructs and how the tests operationalise them in order to generate score/results (i.e. based on the socio-cognitive approach which underpins the CEFR model of language use and which Cambridge English also use in assessment). The operationalisation of this framework led therefore to the development of the online content analysis tool which was then used to examine each exam in detail and to produce the data necessary for the comparability of the tests.

Designing the content analysis tool

Through the above operationalisation of the analytical framework, a questionnaire was designed which included questions addressing each of Kolen and Brennan's four levels of comparability structured around the eight following main sections:

1. Introduction
2. The exam/test: High-level description
    - Design and purpose
3. Goals of language education
4. Speaking:
    - Rating
    - Tasks
5. Writing:
    - Rating
    - Task input/prompt
6. Reading
7. Listening
8. Structural competence

This instrument was reviewed by two experts in language assessment, who provided insightful feedback to make the questions clearer and more precise. The questionnaire was then uploaded onto SurveyMonkey, a secure online survey provider which allows personalisation most features in the questionnaire. To facilitate the work of the experts who conducted the content analysis, one copy of the questionnaire was created for each examination that had to be analysed, allowing the entry of information for all the languages in which each exam was offered.

Content analysts: Selection, training and support

Once the content analysis tool was ready, it was necessary to recruit and train a number of highly qualified experts in language assessment who could feel confident to use this tool to analyse one by one all the language examinations included in the study. These experts were selected from different EU countries on the basis of their expertise in language assessment as well as their knowledge and understanding of relevant foreign languages.

The 16 selected content analysts were requested to virtually attend a webinar where the study was presented to them, their task described and explained, and the content analysis tool was introduced to them together with examples and clarifications. The analysts had the chance to ask questions throughout the webinar and to provide feedback on any difficulties they could anticipate.

After the webinar, a support site was created on Basecamp, an online project management platform which allowed uploading the recording of the webinar in case they wanted to watch some parts of it again, as well as supporting materials which explained in detail how to respond to each question in the online survey. A forum was created within Basecamp for the content analysts to ask any questions they may have while conducting their work. The project team was then able to reply to their questions, and the answers were available to the rest of the content analysts in this forum to reduce the team's workload, as most of the queries regarded similar issues across jurisdictions and examinations.

Allocating work to suitable content analysts: Juggling time and resources

In order to complete the content analysis work within such a short timeframe (hardly five weeks between mid-May and the end of June 2015), it was necessary to ensure that work would be allocated to the analysts in the most efficient way. For this purpose, a number of factors had to be taken into consideration:

- the languages that analysts were proficient in (minimum B2 level required)
- the languages that analysts had a passive understanding of
- the time availability of the analysts.

These three factors had to be combined with the corresponding requirements of the task at hand: the languages tested in the examinations, the language in which each jurisdiction provided the supporting documentation, and the estimated time required to complete the analysis of each examination. Therefore, at this stage the main difficulty was to efficiently juggle work packages, expertise, and time availability to ensure the completion of the task by the given deadline.

In order to meet this deadline, it was decided that the best way to allocate work was asking each analyst to look at the several language versions of each examination. Although the level of proficiency tested may vary across languages, most exams offered in several languages follow very similar procedures and specifications for all the language versions, which significantly facilitated the processing work of the content analysts. Based on this, the main factor considered when allocating work to the content analysts was the languages in which they were proficient (a B2 minimum level was required).

Furthermore, most of the support documentation requested from the countries was provided in the official languages of the countries and there was not enough time to prepare translations of these documents. Instead, content analysts were assigned exams for which they would have an advanced level of the tested languages and, in addition, these exams would have to be from jurisdictions with an official language that the content analyst could understand, at least partially. For example, one of the Polish content analysts who spoke German and English proficiently was assigned jurisdictions such as Slovakia, Slovenia and Czech Republic, where the main language tests are in English and German, and for which the analyst could understand most support documents thanks to the similarity between Polish and the official languages in each of these jurisdictions. This was the second factor which led the allocation of work to the different content analysts.

In some cases, it was not possible to find a totally suitable analyst with both advanced knowledge of all the tested languages and a working understanding of the language in which the support documents were written, especially considering that some jurisdictions were qualified to include up to four languages e.g. Sweden. This issue was overcome by identifying members of staff across the organisation who could understand the language of the given jurisdiction, and arranging for them to support the content analysts as required. This solution proved very successful as content analysts just had to provide a list of relevant information they needed from the support documents, and the native speaker would quickly scan the documents, locate the information requested and translate it into English for the content analyst to record it in the system. In some cases, the native speakers of these languages were also able to provide background and contextual information about the examinations and the language education system in their specific jurisdictions, which proved very helpful for the content analysts to better understand how language tests fit into the wider language education agenda in these given jurisdictions. Table 3 provides an overview of the native and additional languages known by the content analysts.

Table 3: Details of analysts' native and additional languages

Analysts	Native language	Additional languages
Analyst 1	English/French	German
Analyst 2	English	Spanish, German
Analyst 3	Hungarian	English, German
Analyst 4	Italian	English, German
Analyst 5	Dutch	English, Spanish
Analyst 6	Dutch	English, French, German
Analyst 7	Polish	English, German, Swedish, Russian
Analyst 8	Polish	English, Spanish
Analyst 9	German	English, French
Analyst 10	German	English, French, Italian
Analyst 11	English	Spanish, German, French
Analyst 12	Bulgarian	English, Russian
Analyst 13	Italian	English, Swedish, German, French
Analyst 14	Italian	English, German
Analyst 15	Spanish	English, German
Analyst 16	Spanish	English, French

Finally, content analysts' availability and the estimated number of hours required to complete the analysis of the examinations in each jurisdiction also played a key role when allocating work packages to the analysts. Although an initial plan was set out at the very beginning in an effort to anticipate which analyst would be completing each work package, different circumstances ended up heavily affecting the implementation of this plan (i.e. sickness, difficulty to get hold of the required supporting documentation, exam samples not provided in time, etc.). These complications required an extremely close and careful monitoring of the progress, and great flexibility to reshuffle and reallocate work packages to the most suitable analysts to ensure that they had availability to complete the tasks.

To the task: The content analysis work

The content analysis was conducted in order to examine the extent to which test results were likely to be comparable, and it was carried out with the help of the analysis tool presented above. This tool was in the form of an online survey provided on SurveyMonkey, and each content analyst was provided with one copy of the online questionnaire for each examination they had to analyse. The questionnaire allowed the analysts to fill the information for each of the language versions of each examination, so that the background details about the language education context in which the examination existed only had to be entered once.

Content analysts made use of the available data (information provided by the Eurydice Network, example tests and other documents provided by jurisdictions) to answer in as much detail as possible all the questions regarding each of the examinations. The full text of the content analysis tool is included in Appendix 2 of the main report (European Commission 2015). As explained above, content analysts also had access to a specific Basecamp platform, which was used throughout the content analysis stage as a way for content analysts to ask questions regarding the work and to share useful tips and information which could be helpful to other analysts.

Although the content analysts had been picked for this task on the basis of their expertise in language assessment, their work went through a process of spot checking to identify any potential clerical mistakes and misunderstandings. Some minor changes to the data had to be made as a result of this checking. Additionally, for quality assurance purposes, 30% of the examinations were also analysed by a second expert, which ensured the consistency and reliability of the judgements made by the analysts, and allowed for any discrepancies in their understanding of the task to be identified and addressed.

Extracting and cleaning the data for statistical analysis

Once the content analysts had finished filling in the information about the examinations, the data was extracted from SurveyMonkey. Before it could be analysed, the data had to be partially recoded to ensure that responses were compared in a meaningful way, especially in the case of some open field questions which could be all transposed into a common unit e.g. exam duration, length of each exam component, etc. Collation was also necessary to ensure that each language version of each test was considered individually, and that all relevant information regarding the educational background in which this exam is embedded (which was collected in blocks for all the language versions of the same examination, as explained above) had been appropriately copied to each of the language versions.

The descriptive analysis of the data was conducted in Microsoft Excel 2010, and relevant charts and tables were created to illustrate each of the aspects of comparability discussed. In order to establish the most appropriate presentation of the results, three aspects were taken into consideration:

1. **Questions concerning the test construct**: Any questions concerning test constructs and their comparability across examinations were examined separately for each skill. Not all examinations tested all the skills for all the languages, so the results regarding the comparability of constructs need to be considered, bearing in mind which examinations actually tested which skills at which ISCED level.
2. **Questions concerning the CEFR level of the examinations**: There were two contrasting sources for questions regarding the CEFR level of the examinations. On the one hand, test providers and the data facilitated by Eurydice include claims about the intended CEFR level tested by each of the language versions of their examinations. On the other hand, the content analysis tool required analysts to make an expert judgement regarding the difficulty of each task in CEFR terms. This allowed providing an indication of how accurate claims are regarding the tested CEFR levels in each case, and highlighted how poor targeting could jeopardise the comparability of national results across jurisdictions since there would be a mismatch between what candidates should be expected to do at a certain CEFR level and what they can actually do.
3. **Questions not concerning the construct or the CEFR level**: The remaining questions which were not specifically related to the construct or to the CEFR level were grouped either by ISCED level or as a single group. This was done following European Commission/Education, Audiovisual and Culture Executive Agency/Eurydice (2015) indication that exams at ISCED 3 seem to be considered as higher stakes, and therefore it was expected that their characteristics and procedures would differ from those of exams at ISCED 2.

Results: What the study revealed

The qualitative analysis of the language examinations revealed considerable diversity across jurisdictions, which significantly reduces the potential for any meaningful comparisons of test results. The results were analysed according to the operationalisation of the analytical framework presented above, which focused on the following main aspects of comparability:

1. **Inferences (interpretation of results)**: In cases where the tests do not claim alignment to the CEFR, it proved practically impossible to understand how test results had to be interpreted, which hinders any potential for meaningfully comparing these test results across jurisdictions.
2. **Constructs (what is measured by the test)**: Although test components are usually referred to under the same headings across examinations in all jurisdictions (e.g. 'Reading'), the constructs being tested seem to be actually measuring different abilities. This means that two reading tasks from different exams may require a completely different set of competences in the target language, making it therefore very difficult to meaningfully compare the results of tests which are actually measuring different constructs.
3. **Populations (who takes the exams)**: Age was the only demographic characteristic which could be included under this category, and results show that the ages of students taking ISCED 2 and ISCED 3 tests were reasonably homogeneous.
4. **Measurement characteristics/conditions (contextual features which may affect comparability)**: A wide range of measurement characteristics and conditions were considered under this category. In instances where enough evidence was available to make an informed judgement, the diversity across all the measurement characteristics and conditions was considerable. This also raises questions regarding the reliability of the tests, which in turn decreases the potential for comparability of results collected through these tests.

On the basis of the actual findings from the analysis of the language examinations, a number of proposals for *ex-post* adjustment and development work were put forward, as requested by the Tender Specifications. Besides general recommendations for the creation of new language examinations and further developing existing language tests, the most important proposal for *ex-post* adjustment was the application of a new quantitative technique which could overcome the issue encountered by many national examination boards regarding the use of pretesting in their examinations. Jones (this issue) describes in more detail this innovative psychometric technique, its potential to align samples of performance to the CEFR, and how it was piloted with exam task difficulty as part of this study.

Furthermore, the Tender Specifications also required an overview of the data that is currently available from all jurisdictions regarding results of language tests, with the ultimate intention of compiling a European summary table of adjusted results which could then be used to monitor students' progress in language competences. The data that different jurisdictions make available differs greatly, and so does the format in which this information is provided. This presents an added layer of difficulty for comparability purposes, as comparisons of national results are only possible when the data being compared has sufficient elements in common. However, and most importantly, even if the data was produced and reported in a uniform way across jurisdictions, the huge diversity which emerged from the analysis of the examinations, particularly in terms of constructs, inferences and measurement characteristics and conditions, makes it extremely difficult to establish any meaningful comparisons of results from different language examinations. As concluded in the Executive Summary (European Commission 2015:8):

> The meaningful comparability of national results of language examinations across EU Member States will therefore depend not only on these results being expressed in a uniform format, but also on implementing measures at both national and European level that would increase the quality of current language examinations, and in turn ensure that results are similarly valid and reliable across all jurisdictions.

Follow-up: European Commission's new measures to promote multilingualism in Europe

As stated in the introduction, promoting multilingualism is at the heart of the European Union's mission. Languages are key to ensure more inclusive societies, higher employment and growth rates, better conditions for citizens' mobility across MSs, and ultimately a richer understanding of other cultures and lifestyles. Languages are also currently considered as transferable and transversal basic skills, highly needed by the labour markets and with the potential to make both individuals and economies more competitive.

In view of the results from this study, complemented by the report prepared by Eurydice with an overview of all the different language assessment systems in all EU MSs (European Commission/Education, Audiovisual and Culture Executive Agency/Eurydice 2015), the European Commission abandoned the idea of using national data as a way to monitor Europeans' progress in language competences. While they recognise the potential of assessment to improve language education, the results show that many jurisdictions focus on what can be easily tested (e.g. reading skills) rather than on the competences which seem to be required by the labour market (e.g. speaking skills). Furthermore, the Commission is aware of the growing number of international surveys which countries are currently involved in (e.g. Programme for International Student Assessment (PISA), Teaching and Learning International Survey (TALIS), etc.) and they are concerned about the impact that these surveys may have on national education systems, potentially turning them into more test-driven systems rather than encouraging sustainable quality improvements.

In line therefore with the May 2014 Council 'Conclusions on Multilingualism and the Development of Language Competences', the Commission has suggested to take a new set of measures to promote multilingualism and increase the quality of language learning and teaching across EU MSs. These measures include:

- Investing in the improvement of the Online Linguistic Support tool: This tool was designed to test students' language skills at the beginning and end of an Erasmus+ mobility grant, and also includes online language materials aimed at supporting the development of students' language skills while abroad.
- Encouraging a better understanding and use of the CEFR within national education systems: This will mainly be done in co-operation with the Council of Europe and their European Centre for Modern Languages (ECML) through the RELANG project. The main objective of this project is to ensure the correct alignment of national language examinations to the CEFR, which is partly achieved by organising workshops and seminars with language teachers and teacher trainers in the different European countries.
- Emphasising the importance of including effective methods of formative assessment in national education systems: Teachers need to become more familiar with the CEFR and with different methods to incorporate effective formative assessment in language teaching, which will eventually empower them to better determine the language proficiency of their students in CEFR terms which will be comparable across EU MSs.
- Focus on developing students' language-learning skills: Rather than concentrating on teaching students one or two foreign languages, national education systems should seek to develop students' ability to learn and communicate successfully in foreign languages throughout their lives when and as required by their personal and professional circumstances. The world is increasingly globalised and fast-changing and it has become very difficult to predict the needs of the future labour markets. For this reason, the only model of language education which may be successful in a language-rich territory such as Europe is one which will enable citizens to develop language competences in whichever languages they may need throughout their lives. This will ensure that they can fully benefit from the personal and professional opportunities that Europe has to offer, and which will ultimately make Europe a richer and more competitive geographical territory.

References

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

European Commission (2012a) *First European Survey on Language Competences: Final Report*, Luxembourg: Publications Office of the European Union, available online: eceuropaeu/languages/eslc/index.html

European Commission (2012b) *Rethinking Education: Investing in Skills for Better Socio-economic Outcomes*, Luxembourg: Publications Office of the European Union.

European Commission (2014) *Study on Comparability of Language Testing in Europe: Tender Specifications*, Luxembourg: Publications Office of the European Union.

European Commission (2015) *Study on Comparability of Language Testing in Europe: Final Report 2015*, Luxembourg: Publications Office of the European Union.

European Commission/Education, Audiovisual and Culture Executive Agency/Eurydice (2015) *National Tests in Languages in Europe 2014/15*, Luxembourg: Publications Office of the European Union.

Eurostat (2013) *European Day of Languages – Two-thirds of working age adults in the EU28 in 2011 state they know a foreign language. English studied as a foreign language by 94% of upper secondary pupils*, available online: ec.europa.eu/eurostat

Eurostat (2014) *European Day of Languages – English, French and German still most common foreign languages studied at lower secondary level in the EU28 in 2012. . .but Spanish learning has increased more*, available online: ec.europa.eu/eurostat

Kolen, M J and Brennan, R L (2004) *Test Equating, Scaling, and Linking: Methods and Practices* (2nd edition), New York: Springer.

'No More Marking': An online tool for comparative judgement

NEIL JONES CONSULTANT, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Comparability

Comparison is at the heart of meaning: as Laming (2004:9) states, 'there is no absolute judgment. All judgments are comparisons of one thing with another'. Our understanding of the world depends substantially on comparative judgement. Tests designed to measure human competences such as language proficiency can be seen as formal instances of a fundamental human desire to organise experience through classification and comparison. This propensity may be largely unconscious, while formalised measurement requires an explicit design for comparison. This paper discusses a formal psychometric procedure for constructing a measurement scale, derived from comparative judgements.

The No More Marking website is introduced here in the context of a European Commission project (European Commission 2015), which set out to evaluate the possibility of using countries' existing exams and assessments to make valid comparisons of performance. Could these be made to provide accurate and relevant information, perhaps with positive benefit for language education in Europe? The first European Survey on Language Competences (ESLC) presented its findings in 2012, demonstrating wide disparities in countries' language-learning achievements (European Commission 2012a, 2012b). The extent of the difference was such that the Commission's proposal to create a single European benchmark on languages (European Commission 2012c) was rejected by the European Parliament (the rejection also reflected countries' fatigue with international educational surveys). Instead, the Parliament invited the Commission to explore the extent to which existing national systems of data collection regarding language proficiency could be compared across jurisdictions.

The study on comparability of language testing in Europe, undertaken by Cambridge English Language Assessment on behalf of the Commission, set out to address this question. Relatively much smaller in scope than the ESLC, the study was limited in terms of its timeframe and its access to data.

The study combined qualitative and quantitative approaches: a structured descriptive survey of countries' assessments and tests, conducted by suitable experts (see Saville and Gutierrez Eugenio, this issue), combined with an attempt to locate countries' language-learning achievements on a measurement scale, criterion referenced to the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001). Not unexpectedly, the qualitative analysis identified a range of factors differentiating countries' language educational goals and their practice of assessment, suggesting that comparability would be difficult to achieve. The quantitative project described here attempted to draw some kind of straight line through what was clearly a very diverse reality.

In the event, the study was somewhat limited by the project's timeframe and the data that countries could provide: in fact, the process of collating and interpreting data for the study illustrated an additional threat to the notion of comparability. The original aim was to locate countries' test tasks on a scale of difficulty, and then use these to estimate the levels of ability demonstrated by learners taking these tests. That second stage was not achieved, although an illustration of the concept was provided – see the section 'Interpreting national/regional levels of performance' in European Commission (2015:153).

Comparative judgement

Comparative judgement (CJ) refers to the organised collection and analysis of judgements about entities such as candidates, test forms, flavours of ice-cream, or anything where human judgement is of value. What makes it comparative is that the judgements are *relative*, rather than absolute. Judges are asked to make a series of judgements as to which of two entities is higher, harder or tastier. The final outcome is a set of entities ranked on a scale from lower to higher. The more judgements that contribute to this ranking, of course, the more precise the outcomes will be.

For the background to CJ, see Thurstone (1927), Andrich (1978), Bramley (2005) and Jones (2009). CJ exploits similar insights to the Rasch model: that a measurement scale can be

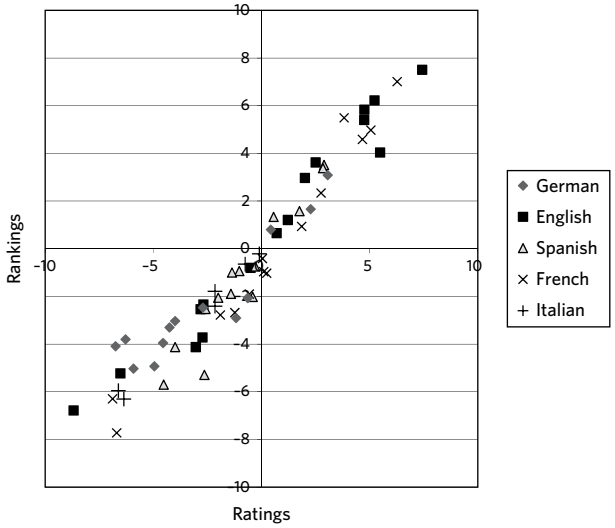
constructed from the interactions of persons and test items, where personal abilities can be abstracted from task difficulties (the notion of additive conjoint measurement). CJ can be located within the wider framework of Item Response Theory (IRT) (Drasgow and Hulin 1990, Embretson and Reise 2013).

One issue discussed by Bramley and others using paired comparisons is the large number of paired judgements required. A ranking approach to CJ, where more than two objects are simultaneously compared, is discussed by Bramley (2005) and Linacre (2006).

Such a ranking approach to CJ was used for a multilingual benchmarking conference organised by Centre international d'études pédagogiques (CIEP) at Sèvres in June 2008, which focused on the performance skill of speaking (Breton 2008). It is interesting to report on here not simply because it relates to language proficiency, but because two kinds of data were collected, based on judgements of video samples: CEFR criterion-related judgements during the conference itself, but prior to the conference, ranking data collected from the same judges, using an online platform that allowed them to record their ranking by dragging samples within a list.

Figure 1 compares the abilities estimated from rankings and ratings for the set of samples submitted to both procedures. The correlation is high (0.94), particularly given the absence of rater training in advance of the conference.

Figure 1: Ranking and rating compared (Breton 2008)



Criterion-referenced CEFR levels were assigned according to the judgements made at the rating conference. These, it was felt, could be treated as definitive.

At the same time the event provided an impressive validation of the CJ approach. This example demonstrated that the data provided by a CJ event is amenable to analysis using the Rasch model. The rating and ranking data are comparable, as are the measurement scales which they enable us to construct.

The CIEP study also indicates how criterion-related standards may be applied to the CJ data, given that in CJ judges are not asked to identify a level of performance in absolute terms, e.g. as representing a borderline B1 level performance on the CEFR. In the CIEP study it was possible to base the criterion-referenced levels on the joint expertise of the conference members with regard to the aligned languages. But the same data could be used to align further languages without repeating the standard setting – simply by conducting a further comparative study. If examples of absolute standards already exist in the dataset then the CJ procedure will automatically align performances to their correct relative position in the ranked dataset. In the case of the comparative study reported in this paper, tasks taken from the ESLC were used as anchors in this way.

'No More Marking'

Comparative judgement thus provides simple ways of bringing psychometric procedures to bear on organising and standardising human judgement in order to play to its strengths – that is, by making relative rather than absolute decisions. While the principles of CJ have long been understood, and specific approaches to analysis within the IRT paradigm have been developed, there seems to be a new interest in exploiting CJ to address assessment issues. No More Marking is the name of the website which hosted the data and the analysis for the study on comparability of language testing in Europe. As the name suggests, this site is targeting an interest in using CJ to improve upon traditional approaches to marking used by assessment bodies engaged in educational testing. An instructive interactive report available on the website illustrates the results of a 2015 Ofqual study conducted through the No More Marking website, in which Maths PhD students judged maths items from a range of sources, addressing the question: 'Which question is the more mathematically difficult to answer fully?' In addition to the interactive report the study is presented in Ofqual (2015). It confirms the power of CJ to compare standards and reveal issues in a way which would otherwise be impossible. It will be very interesting to see how exam boards respond to these new technology-based opportunities.

The orientation of the No More Marking website is clear: the objects of comparison are named 'candidates', although that is no obstacle to using the system for other entities – test tasks, in the case of our study. There is also no problem in working with several different objects of comparison. In our study there were six datasets: French, English and a bilingual French/English anchor, and one for each of these two skills: reading and writing.

Some findings from CJ data in the study on comparability

This section is not intended to provide a complete account of the survey's findings, which are available in European Commission (2015). The goal is to illustrate the use of CJ and the potential of the No More Marking platform, not only for identifying interesting patterns in the data, but also for validating the judgements made and providing indices of quality for the outcomes.

Much of the description and analysis based on CJ data focused on the CEFR levels and how countries' test tasks related to them by level of task difficulty. The apparent

difference in difficulty between the two ISCED levels included in the study is illustrated by Figure 2. One would expect that the difference between ISCED 2 and 3 would be evident, if countries successfully distinguish the levels in their test designs. Concerning countries' tests of English this is the case: ISCED 2 is seen to be about B1–B2, ISCED 3 about C1.

Additionally, this methodology has the potential to show the location of individual countries' test tasks; however in the case of this study the samples were quite small so the precision of such estimates was not high.

For French, however, the picture was not as clear, as shown in Figure 3.

Figure 2: English reading and writing: ISCED levels

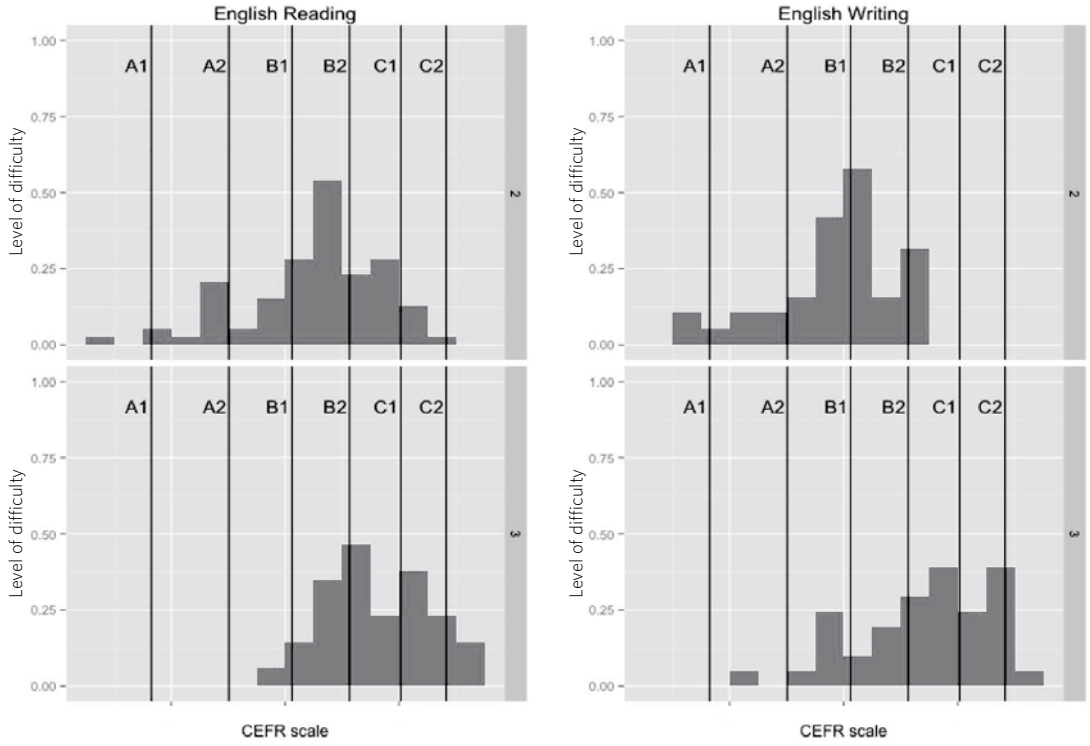
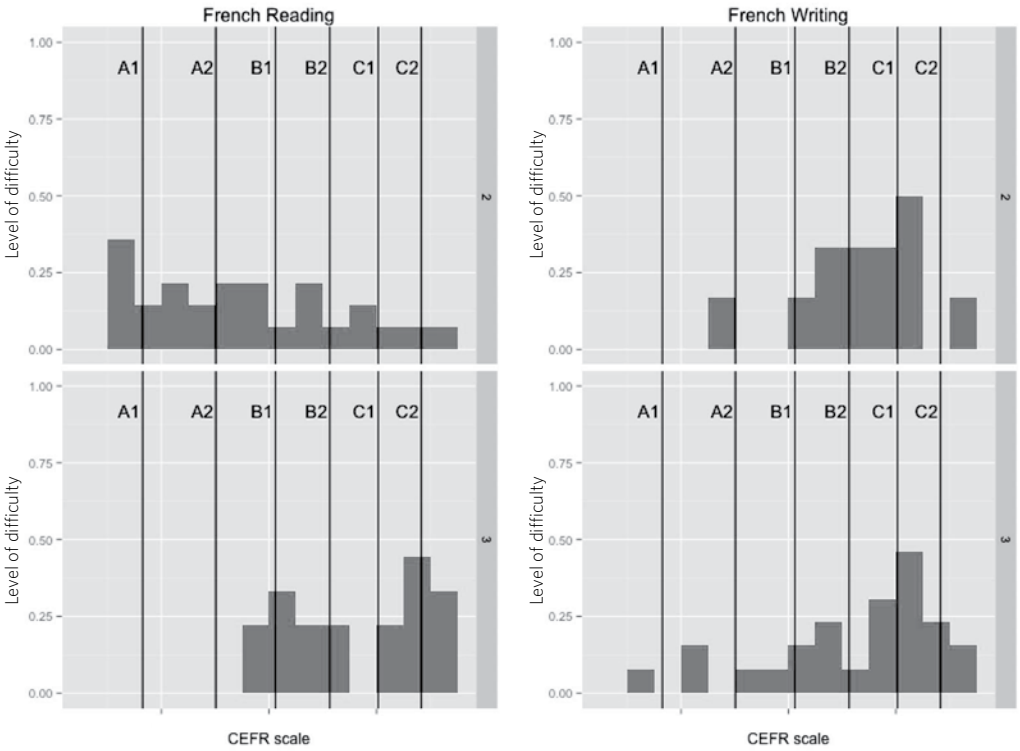


Figure 3: French reading and writing: ISCED levels



Was this a true result or was there an issue with the data? Certainly the process adopted to put French and English on the same scale was quite complex, involving a bilingual French–English anchor in which judges with a competence in both languages ranked them together (replicating the CJ procedure in the CIEP study). The intention was to take English as the fixed point and equate French to it. Thus English was equated to the bilingual set, and the bilingual set then equated to French.

This was a different approach to that adopted for standard setting in the ESLC, which was based on the judgements of language-specific expert groups. By using the resources of the CJ platform we hoped to validate an approach to standard setting which did not depend on this requirement.

Studying rater performance

A valuable feature of No More Marking is the analysis of rater performance. This includes an IRT *infit* statistic which evaluates the relative consistency of rater judgements: the more raters agree with each other the greater our confidence in their findings.

Studying the infit for the datasets identified above we found interesting effects. There was good fit in the case of English and French (monolingual) reading, but relatively large misfit in the case of the bilingual English/French reading dataset. This may indicate that the task of judging the relative difficulty of test tasks in different languages is actually more difficult than we appreciated. However, as the infit statistic is sensitive to the number of judgements made and it transpires that far fewer judgements were available for the French tasks, we cannot come to firm conclusions from this data.

Conclusions

Comparative judgement is potentially a powerful tool which can provide insights when other approaches are not applicable. Like any tool, it needs careful use and its limitations must be considered. Sufficient raters must be recruited to achieve the degree of reliability required; some approach to training is important, to ensure that raters are using the same set of criteria. The analysis of rater performance is valuable, and studying it early may be important in identifying problem raters.

This article has presented evidence for the similarity of outcomes from CJ and more traditional rating, and arguably this is an area deserving closer study. The use of CJ to construct a measurement scale where no scale previously existed is straightforward (and valuable); but where the aim is to relate to an existing scale, such as one linked to the CEFR levels, then one cannot take for granted that the two scales will exhibit a perfectly linear relationship. Different approaches to scale construction can produce quite different results – for example, the CEFR descriptor scales developed by North (2006), based on teachers' judgements of Can Do statements, give quite a different picture to that provided by the scale of CEFR-related exam levels developed by Cambridge English, which is based on anchoring estimates of ability derived from tests over a series of levels. Comparative judgement may also have its own specific effects.

Today in most parts of Europe at least, most educational testing or assessment does not make use of strong psychometric models based on IRT, so that although governments may insist on the importance of 'standards', there are only weak mechanisms in place for determining whether standards are rising, falling or staying much the same. One reason for neglecting the strong psychometric technologies which offer a grasp on standards is that they involve pretesting, in order to determine the difficulty of tasks, and pretesting is widely seen as impossible for reasons of test security. Thus a valuable use of CJ in assessment might be to provide data for an IRT approach which avoids pretesting, by enabling this year's tasks to be equated to last year's tasks securely by a panel of judges. Might we ask: No More Pretesting?

References

Andrich, D (1978) Relationships between the Thurstone and Rasch approaches to item scaling, *Applied Psychological Measurement* 2, 449–460.

Bramley, T (2005) A rank-ordering method for equating tests by expert judgment, *Journal of Applied Measurement* 6 (2), 202–223.

Breton, G (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)*, le centre international d'études pédagogiques, Sèvres, 23–25 June 2008.

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

Drasgow, F and Hulin, C L (1990) Item response theory, *Handbook of Industrial and Organizational Psychology* 1, 577–636.

Embretson, S E and Reise, S P (2013) *Item Response Theory*, Abingdon: Psychology Press.

European Commission (2012a) *First European Survey on Language Competences: Final Report*, Luxembourg: Publications Office of the European Union.

European Commission (2012b) *First European Survey on Language Competences: Technical Report*, Luxembourg: Publications Office of the European Union.

European Commission (2012c) *Rethinking Education*, available online: ec.europa.eu/languages/policy/strategic-framework/rethinking-education_en.htm

European Commission (2015) *Study on Comparability of Language Testing in Europe: Final Report 2015*, Luxembourg: Publications Office of the European Union.

Jones, N (2009) A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, *Research Notes* 37, 6–9.

Laming, D (2004) *Human Judgment: The Eye of the Beholder*, London: Thomson Learning.

Linacre, J M (2006) Rasch analysis of rank-ordered data, *Journal of Applied Measurement* 7 (11), 129–139.

North, B (2006) *The Common European Framework of Reference: Development, Theoretical and Practical Issues*, paper presented at the symposium 'A New Direction in Foreign Language Education: The Potential of the Common European Framework of Reference for Languages', Osaka University of Foreign Studies, Japan, March 2006.

Ofqual (2015) *GCSE Maths, Final Decisions*, available online: ofqual.blog.gov.uk/2015/05/21/gcse-maths-final-decisions

Thurstone, L L (1927) A law of comparative judgment, *Psychological Review* 3, 273–286.

Updating the CEFR descriptors: The context

BRIAN NORTH EUROCENTRES FOUNDATION, SWITZERLAND
JOHANNA PANTHIER LANGUAGE POLICY UNIT, COUNCIL OF EUROPE

The background to the CEFR

The *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR) was designed to provide a transparent, coherent and comprehensive basis for the elaboration of language syllabuses and curriculum guidelines, the design of teaching and learning materials, and the assessment of foreign language proficiency. It was published by the Council of Europe in 2001 in English and French (Council of Europe 2001a, 2001b). Today the book is available in 40 different languages, including non-European languages such as Arabic, Chinese, Japanese or Korean and is widely used as a reference tool for curriculum development, teacher training, pedagogical material and assessment of language competences in Europe and beyond (Byram and Parmenter (Eds) 2012), but many users are not aware of its origins or context. It is often assumed that it is a document produced by the European Union (EU).

The Council of Europe and the EU are two different organisations. The Council of Europe was founded in Strasbourg in 1949 with 10 members, as one of the measures taken immediately after World War II in order to promote democracy, protect human rights and the rule of law in Europe. The EU has its origins in the European Coal and Steel Community, founded in 1951 by the Treaty of Paris; the Treaty created a common market and evolved through subsequent treaties to become the EU, which currently has 28 members. No country has ever joined the EU without first belonging to the Council of Europe. The two organisations have their specific aims and objectives, but share common values and co-operate closely together.

The Council's commitment to its fundamental values underpins its intergovernmental co-operation programmes in the field of education. The 47 current member states work together in a process of mutual support to ensure the right to education and an education of quality for all, which is an essential foundation for social inclusion and social cohesion.

The right to education is enshrined in the European Convention on Human Rights and Fundamental Freedoms (Additional Protocol No.1, Article 2). Further support for the implementation of this basic right is provided by the European Social Charter which guarantees social, economic and cultural rights, including the right to accessible and effective education. This human rights perspective is further promoted in Recommendations from the different organs of the Council of Europe, in particular the Committee of Ministers. These non-binding policy texts offer a broad European consensus on guiding principles and possible measures to ensure enjoyment of the right to education, as set out, for example, in: Recommendation (2012/13) on ensuring quality education; Recommendation (2014/5) on 'the importance of competence in the language(s) of schooling for equity and quality in education and for educational success', which stresses that the right to education can only be fully exercised if learners master the specific linguistic competences that are necessary for access to knowledge.

In a similar vein Recommendation (2008/7) on the use of the CEFR and the promotion of plurilingualism invites the member states, inter alia, '. . . to create and/or maintain conditions favourable to the use of the CEFR as a tool for coherent, transparent and effective plurilingual education in such a way as to promote democratic citizenship.' *Plurilingualism* is the Council of Europe's response to European linguistic and cultural diversity. Plurilingualism differs from *multilingualism*: plurilingualism is the ability to use functionally more than one language – and accordingly sees languages from the standpoint of speakers and learners. Multilingualism, on the other hand, refers to the presence of several languages in a given geographical area, regardless of those who speak them. A given society can be multilingual by the co-existence of monolingual citizens speaking each a different language, or by the presence of plurilingual citizens. The Council of Europe has clearly opted for plurilingualism, as competences in different languages open doors to and respect for other cultures and allow citizens to pursue the learning of specific languages according to their evolving needs. The Recommendation recognises the potential of the CEFR 'for the diversification of language learning within educational systems in order to maintain and develop plurilingualism among citizens of Europe as a means of knowledge building and skills development, with a view to enhancing social cohesion and intercultural understanding'. The CEFR (also the related European Language Portfolio) has been adapted for use in a variety of contexts characterised by a strong human right and social inclusion dimension, for example, to support the language learning of migrant children, adolescents and adults, as well as for learning minority languages, including Romani. The evolution of the use of the CEFR, developed and supported by the Language Policy Unit of the Council of Europe in Strasbourg, is therefore twofold. It needs to serve not only as a basic reference tool for foreign and second language education, but also as a flexible and dynamic reference instrument that can be adapted to context in a manner that contributes to the promotion of the Council of Europe's core values and its overarching goal of plurilingual and intercultural education. As outlined in the *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education* (Beacco, Byram, Cavalli, Coste, Egli Cuenat, Goullier and Panthier 2015), the purpose of plurilingual education is to establish areas of convergence in the teaching of different languages; intercultural education aims at developing the ability to experience otherness and diversity, analyse that experience and derive benefit from it. Current CEFR developments can also be considered against this background in terms of their particular added value to the mission of the Council of Europe and its education programme.

The Council of Europe's involvement in the language area followed the signing of the European Cultural Convention in 1954, and started in earnest in 1960. The main recommendation resulting from the Council of Europe's 1964–1973 major project concerning modern languages was a European-wide unit/credit scheme for adult learners of modern languages. An intergovernmental symposium took place in 1971 in Rüschlikon, Switzerland, to develop the scheme. It focused on three areas: (a) new forms of organisation of linguistic content; (b) types of evaluation within a unit/credit scheme; and (c) means of implementation of a unit/credit scheme in the teaching/learning of modern languages in adult education. At the Symposium, John Trim and David Wilkins, the UK delegates, sketched out the *action-oriented approach* that became the basis of *The Threshold Level* (van Ek 1976, van Ek and Trim 2001b) and of the CEFR. 'Threshold' is a notional/functional specification of the language knowledge and skills needed to visit or live in another country or, more generally, to communicate in an independent way with people who speak the target language. It provides lists of relevant situations and texts plus a detailed analysis of the general notions (like space, time, possibility, probability), the specific notions (more akin to topics) and the language functions that people will need in such situations, providing appropriate language exponents, together with an analysis of the requisite syntactic, morphological and phonological content operationalised in them.

Threshold-level specifications are available for nearly 30 languages. The *action-oriented approach* of the Threshold Level considers the learner as a *language user*. Language learning is not perceived as an intellectual pursuit to train minds, but as a practical skill to communicate with others. This approach is central to the CEFR. Can Do descriptors were also discussed in 1971; they were familiar from the US Foreign Service Institute (FSI) language proficiency scale, published by Wilds in 1975. The Council of Europe's project and the Threshold Level specifications had an immense influence on curriculum, course books and examinations. Can Do descriptors were developed for self-assessment of various aspects of Threshold content by Oscarson (1979, 1984). A cut-down version of Threshold representing specifications for a stage approximately halfway towards it (appropriately called *Waystage*) soon followed, to be joined in 1990 by an expanded version called *Vantage*. Therefore, by the time work on the CEFR started, there was an emerging set of Council of Europe common reference levels, although these were defined in terms of content specifications rather than performance standards. Based on this work, Eurocentres, for example, developed a common Can Do descriptor scale of language proficiency for English, French, German, Spanish and Italian during the 1980s, with content specifications and criteria for assessment at each level. Cambridge ESOL, now Cambridge English Language Assessment, developed examinations for Waystage (*Key English Test*, now known as *Cambridge English: Key*) and Threshold (*Preliminary English Test*, now known as *Cambridge English: Preliminary*) as well as creating the *Certificate in Advanced English* (now known as *Cambridge English: Advanced*) at a level that was later to become C1, filling the gap between *First Certificate in English* (reflected in Vantage and now known as *Cambridge English: First*) and *Certificate of Proficiency in English*, now known as *Cambridge English: Proficiency* (C2).

The CEFR project

The first attempt to move towards a common European framework of objectives had been made in the late 1970s. A proposal for a set of European levels was made in outline by Wilkins (1978) at an intergovernmental symposium in Ludwigshaven, Germany, called to discuss a concrete proposal for the unit/credit scheme for language learning in Europe that had been recommended by the 1971 Rüschlikon Symposium (Trim 1978). That symposium rejected the scheme and recommended instead a series of workshop-projects in order to introduce communicative language teaching to the state school sector. Such workshops ran through the 1980s and formed the basis for the foundation in 1994 of the European Centre for Modern Languages (ECML), which has since continued this work. In 1989, rather than running another workshop project, the Swiss authorities proposed another intergovernmental symposium to attempt to move towards a common framework. This symposium took place at Rüschlikon in November 1991 (Council of Europe 1992) and recommended the development of a Common European Framework of objectives around a set of European language levels. The idea of such common reference levels had formed part of the presentation of the European Language Portfolio at the symposium (Schärer and North (Eds) 1992). Following the symposium, an ad hoc expert group was charged with developing an approach towards such a framework (North, Page, Porcher, Schneider and Van Ek 1992). The report included a concrete proposal for six reference levels based on the Waystage–Threshold–Vantage series of objectives (van Ek and Trim 2001a, 2001b, 2001c). This proposal was inspired (a) by a presentation in a round table discussion at the Rüschlikon Symposium by Peter Hargreaves, then CEO of UCLES (now Cambridge English Language Assessment), and (b) by a desire to maintain coherence with the levels being adopted by the newly founded, Cambridge-coordinated, Association of Language Testers in Europe (ALTE), then being set up with support from an EU Lingua funding programme.

The first versions of the CEFR were then produced between 1994 and 1996 by an Authoring Group consisting of John Trim, Daniel Coste, Joseph Sheils and Brian North, under the supervision of a wider Working Party of European experts. Parallel to this development, a 1993–1996 Swiss National Science Foundation project, using a methodology proposed by North (1993a), developed the basis for the CEFR's illustrative descriptors (North 1995, 2000, North and Schneider 1998, Schneider and North 2000) as well as producing the prototype for the European Language Portfolio on the basis of those descriptors (Schneider, North and Koch 2000). After a piloting phase launched in 1997, the final version of the CEFR was then published in English and French in 2001, the European Year of Languages. The publication of the CEFR itself was followed by the development over the next few years of a CEFR 'toolkit':

- specifications for different languages, called 'Reference Levels' (e.g. French: Beacco, Porquier and Bouquet 2004,

2007, Beacco, De Ferrari, Lhote and Tagliante 2006, Beacco and Porquier 2008; German: Glaboniat, Müller, Rusch, Schmitz and Wertenschlag 2005; Spanish: Instituto Cervantes 2007; Italian: Parizzi and Spinelli 2009; English: the ongoing English Profile; see www.englishprofile.org; more recently for English: North, Ortega and Sheehan 2010; for French: Evaluation and Accreditation of Quality in Language Services and Centre international d'études pédagogiques 2015, see also www.eaquals.org)

- case studies of implementation (Alderson (Ed) 2002)
- illustrative video samples scientifically calibrated to the levels and descriptors for English (Eurocentres and Migros Club Schools 2004), French (Centre international d'études pédagogiques and Eurocentres 2005), German (Bolton, Glaboniat, Lorenz, Müller, Perlmann-Balme and Steiner 2008), Italian (Centro Valutazione Certificazione Linguistica, Perugia, 2006) and for five languages (Centre international d'études pédagogiques 2008)[1]
- illustrative test tasks and items (Figueras and Takala 2016)
- a Manual to help relate tests and examinations to the CEFR (Council of Europe 2003, 2009a) together with further materials offering alternative standard-setting methodologies (North and Jones 2009), a set of related case studies (Martyniuk (Ed) 2010) and a Reference Supplement (Council of Europe 2009b)
- a Manual for Language test development and examining for use with the CEFR – produced by ALTE on behalf of the Language Policy Unit (Council of Europe 2011).

Despite the provision of the Manuals, however, the CEFR is not a standardisation project, as was emphasised at the Language Policy Forum called to take stock of its implementation (Council of Europe 2007). The CEFR merely proposes common reference levels (A1–C2), which can be exploited as reference points in developing *locally appropriate* standards in order to increase transparency and coherence, both for local end users and for professional colleagues elsewhere. The aim of the CEFR in general, and of its Can Do descriptors in particular, was in fact to stimulate educational reform by re-orienting language teaching and learning to real-life needs. The CEFR sees learners as social agents. Few people need to speak a foreign language perfectly and many people need to speak more than one foreign language at least partially. Thus, the concepts of needs analysis, partial competences and plurilingualism introduced with the CEFR are intimately inter-related. This action-oriented perspective has sometimes been misinterpreted as an instrumental, commercial perspective: a capitulation of language education to language training for a product-oriented society. Nothing could be further from the truth, as is recognised by the many contributors to an international survey on the use of the CEFR (Byram and Parmenter (Eds) 2012), particularly Porto (2012). The CEFR was intended, on the contrary, as a political instrument promoting the European dream of creating an informed, plurilingual population able to accept otherness, to empathise and to communicate across linguistic and cultural barriers.

Innovative aspects of the CEFR

There are several aspects of the CEFR that remain innovative, even 20 years after the first version was written:

- the replacement of Lado's (1961) four skills (listening, speaking, reading, writing) with a further development of North's (1992) proposal of the communicative language activities reception, interaction and production, with North's fourth category 'processing' being developed, at Daniel Coste's suggestion, into mediation
- the replacement of Lado's three elements (grammar, vocabulary, pronunciation) with a further development of Van Ek's (1986) model of communicative competence into three communicative language competences: linguistic, socio-linguistic, and pragmatic (discourse and functional)
- illustrative descriptor scales for communicative language activities that produce types of discourse, each of which would have its own socio-pragmatic and discourse conventions
- a positive interpretation of communicative language strategies following Tarone (1983) and Færch and Kasper (Eds) (1983) with descriptor scales for reception, interaction and production strategies, in a context in which strategies had up until then been seen only as compensatory
- an emphasis on the co-construction of discourse in interaction (e.g. see Council of Europe 2001a:99)
- the introduction of the concepts of plurilingual and pluricultural competences (CEFR Section 6.1.3, Council of Europe 2001a:133–135)
- curriculum design which takes into account linguistic diversification, partial competences, plurilingualism and pluriculturalism, differentiated learning objectives and life-long language learning.

With these proposals, the CEFR moves away from the traditional focus in language learning, teaching and testing on linguistic expression. It promotes the notion of collaborative co-construction of meaning to accomplish a given task under the prevalent conditions and constraints, mobilising competences (including plurilingual ones) and positive strategies to do so (see CEFR Section 2.1 (Council of Europe 2001a:9–16) for the approach adopted). Many applied linguists had pointed out that the four skills model did not reflect real-life language use (Alderson and Urquhart 1984:227, Breen and Candlin 1980:92, Brumfit 1984:69–70, 1987:26, Stern 1983:347, Swales 1990; see also North 2000:103–105 for a discussion), but the CEFR was probably the first reference document to replace it. However, inspiration for the CEFR's descriptive scheme came from education as well as linguistics, inspired by pioneering work on collaboration in small groups in the classroom (Barnes and Todd 1977, Brown, Anderson, Shillcock and Yule 1984, Oxford Certificate of Educational Achievement 1984, North 1991, 1993b). These innovative aspects of the CEFR are often missed because, in a work that is more a thesaurus than a thesis, they are mentioned but not

[1] DVDs are available at the Language Policy Unit, Council of Europe, Strasbourg and the recordings can be accessed online: www.ciep.fr/en/books-and-cd-roms-dealing-with-assessment-and-certifications/dvd-spoken-performances-illustrating-the-6-levels-of-the-common-european-framework-of-reference-1

developed. The CEFR tends to list rather than develop aspects that it invites users to reflect on.

In this respect, it is perhaps worth citing how plurilingual and pluricultural competence is discussed in the CEFR. After introducing it as a necessarily uneven, changing and differentiated competence, the text mentions that a plurilingual person is able to alternate between and blend the languages in their repertoire – i.e. to code switch:

> A further characteristic of plurilingual and pluricultural competence is that it does not consist of the simple addition of monolingual competences but permits combinations and alternations of different kinds. It is possible to code switch during the message, to resort to bilingual forms of speech. A single, richer repertoire of this kind thus allows choice concerning strategies for task accomplishment, drawing where appropriate on an interlinguistic variation and language switching (CEFR Section 6.1.3.2, Council of Europe 2001:133–134).

Plurilingual and intercultural education for democratic citizenship

It is this spirit that has inspired the Council of Europe's involvement in the language field since the publication of the CEFR. As the work of the Council of Europe in the language field has developed from concerning adults to concerning schoolchildren, from supporting foreign language education to supporting education in all languages of schooling, and from focusing on language competences to promoting plurilingual and intercultural education, it is these innovative, transversal aspects of the CEFR that have been further developed.

Already in 1999, during a conference entitled 'Linguistic diversity for democratic citizenship in Europe', the form and content of a guide for the development of language education policies were conceived. After two preliminary, pilot versions, the guide entitled *Guide for the Development of Language Education Policies in Europe: From Linguistic Diversity to Plurilingual Education* (Beacco and Byram 2007) was published in two versions (main version and executive summary) in 2007. The aim of this guide is to offer an analytical instrument which can serve as a reference document for the formulation or reorganisation of language teaching in member states. As stated in the Introduction to the Guide:

> The Council of Europe and its member States have taken the position that it is the promotion of linguistic diversity which should be pursued in language education policy. For in addition to mobility, intercomprehension and economic development, there is the further important aim of maintaining the European cultural heritage, of which linguistic diversity is a significant constituent. Thus it is a question not only of developing or protecting languages but equally of enabling European citizens to develop their linguistic abilities. This means, then, that language teaching must be seen as the development of a unique individual linguistic competence ("knowing" languages whichever they may be). This competence needs to be developed not just for utilitarian or professional reasons but also as education for respect for the languages of others and linguistic diversity (Beacco and Byram 2007:7).

The Guide constitutes one of the key documents for the development of (national or regional) 'Language Education Policy Profiles'. The Language Policy Unit offers expertise to assist member States who so wish in reflecting upon their language education policy. Developing such a Profile provides member States, regions and cities with the opportunity to undertake a self-evaluation of their policy in a spirit of dialogue with Council of Europe experts, with a view to focusing on possible future policy developments within the country. So far, 17 'Language Education Policy Profiles' have been finalised and published on the Council of Europe website (www.coe.int/lang) and two others are currently underway.

More recently the Council of Europe has published a *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education* (Beacco, Byram et al 2015). As explained in the Foreword of the document, the decision to develop the Guide was taken at a Council of Europe Policy Forum on challenges and responsibilities in the use of the CEFR (Strasbourg, 6–8 February 2007), because:

> The discussion and exchange at that forum certainly showed beyond question that the CEFR had succeeded at European level. But they also showed that the uses made of it tapped only part of its considerable potential and even, in some cases, disregarded certain values which the Council's member states promote, and which underlie the approaches it describes. This obvious imbalance in implementation of the CEFR's provisions chiefly affects plurilingual and intercultural education, although this is one of the CEFR's main emphases. In fact, few language curricula are consistently geared to such education. Participants at the forum stressed the need for a document which would expound the various aspects of that dimension and explain how it could be implemented, taking as a basis the CEFR and other Council of Europe texts, particularly the *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education* (Beacco, Byram et al 2015:5).

Also in 2015, in the context of its major project on the language of schooling, the Language Policy Unit published *The Place of Languages of Schooling in Curricula* (Beacco, Coste, Linneweber-Lammerskitten, Pieper, van de Ven and Vollmer 2015). This handbook was written to support the implementation of the principles and measures set out in a 2014 Recommendation from the Committee of Ministers of the Council of Europe to its member states concerning 'The importance of competences on the language(s) of schooling for equity and quality in education and for educational success'. It aims to show why language in all subjects is important, and what the implications are for policy and practice.

A rich variety of studies and conference papers arising from the project on the language of schooling are available on the *Platform of Resources and References for Plurilingual and Intercultural Education* (www.coe.int/lang-platform), including procedures for defining the language dimension in curricula for history (Beacco 2010), for mathematics (Linneweber-Lammerskitten 2010), for sciences (Vollmer 2010) and for literature (Pieper 2010). There is also a more general text on the linguistic dimensions of knowledge building entitled *The Language Dimensions in All Subjects: A Handbook for Curriculum Development and Teacher Training* (Beacco, Fleming, Gouiller, Thürmann and Vollmer 2010). All these guides and tools were developed as a response to the major aims of social inclusion and cohesion as defined by the heads of governments of the Council of Europe member states in 2005. They are useful instruments for decision makers wishing to review their language education policy, who can use them independently or ask for a policy review by Council of Europe experts. The documents are freely available on the Council of Europe

website and also address other stakeholders: curriculum developers, authors of pedagogical material, teacher trainers, teachers, examiners, etc.

In addition to these guides and resources related to plurilingual and intercultural education, and to the language of schooling, the Language Policy Unit has also set up another important project concerning the Linguistic Integration of Adult Migrants (LIAM). This is because the integration of migrants and the impact on it of their acquisition of competence in the language(s) of the host country are a focus for political debate and policy initiatives in a growing number of Council of Europe member states. The website of this project (www.coe.int/en/web/lang-migrants) offers a practical means of pooling and accessing many useful resources.

Most recently, the Language Policy Unit also contributes to a transversal project of the Education Department of the Council of Europe concerning Competences for Democratic Culture (CDC). This project has developed a set of descriptors that will eventually be incorporated in a *Framework of Reference of Competences for Democratic Culture: Teaching, Learning and Assessment*.

Updating CEFR descriptors

The development of the additional CEFR illustrative descriptors should be understood in the wider context of the projects mentioned in the previous section. At a meeting in May 2013 the Council of Europe's Education Department decided to commission a text to situate the CEFR's descriptive scheme within this broader educational context for language learning, teaching and assessment that has developed over the past 20 years, and in particular to develop the concept of mediation in an educational setting (Coste and Cavalli 2015). According to their definition (Coste and Cavalli 2015:27), 'mediation can be defined as any procedure, arrangement or action designed in a given social context to reduce the distance between two (or more) poles of otherness between which there is tension'. The Education Department also decided to commission the expanded set of CEFR illustrative descriptors from Eurocentres[2] under the co-ordination of Brian North, developer of the original set (North 1995, 2000). The project, which is just coming to an end, took place in two phases.

Phase 1: 2013–2014 – Reviewing the 2001 CEFR descriptor scales

This phase concerned the exploitation of existing materials to develop additional descriptors for the existing set of illustrative descriptor scales, to complement those available in the CEFR. The focus was on plugging gaps in the original set and further developing the description at A1 and the C levels. The work was undertaken by a small Eurocentres Authoring Group[3] supported by a Sounding Board[4], and reviewed by a wider group of experts in later 2014.[5] The source material was CEFR-related descriptors that had been validated and calibrated to CEFR levels with a similar methodology to that used to create the original set. The projects whose descriptors were exploited were the following:

- ALTE Can Do statements, 2001
- AMMKIA project (Finland: Sauli Takala)
- Cambridge Assessment Scales for Speaking, Common Scale for Speaking, Common Scale for Writing, *BULATS* Global scale (www.cambridgeenglish.org/exams/bulats)
- CEFR-J project for Japanese secondary school learners of English, 2011 (www.tufs.ac.jp)
- English Profile descriptors for the C levels, published in Green (2012)
- Lingualevel/IEF (Swiss) project for 13–15 year olds, 2009: www.lingualevel.ch
- Pearson Global Scale of English (GSE), 2012/2014: www.english.com/gse

In addition, some 50 descriptors from non-calibrated sources were included after calibration by Pearson Education in conjunction with their GSE project.

Phase 2: 2014–2016 – Focus on mediation

At the end of 2013 it was decided to go a step further and develop from scratch descriptors for different aspects of mediation, for which descriptors had not been included in the CEFR in 2001. As a result a Mediation Working Group[6] has, since January 2014, been refining and validating a set of illustrative descriptors from a range of sources, as described in the next article. The approach taken was much influenced by the broader definition of mediation being developed by Coste and Cavalli (2015). Unlike with the four skills model, or its reworking in the CEFR into reception, interaction and production, mediation is not concerned with the linguistic expression of a speaker. Instead, the focus is on the role of language in processes like creating the space and conditions for communication and/or learning, collaborating to construct new meaning, encouraging others to construct or understand new meaning, passing on new information in an appropriate form, and simplifying, explaining, elaborating or otherwise adapting input in order to facilitate these processes (mediation strategies). The context can be social (e.g. Wall and Dunne 2012), pedagogic (e.g. Mercer and Hodgkinson (Eds) 2008), cultural (e.g. Zarate, Gohard-Radenkovic and Lussier 2004), linguistic (e.g. Stathopoulou 2015) or professional (e.g. Lüdi 2014). Coste and Cavalli (2015) propose a fundamental distinction between *cognitive mediation*, the process of facilitating access to knowledge and concepts, particularly when an individual may be unable to access this directly on their own, and *relational mediation*, the process of establishing and managing interpersonal relationships, usually in order to improve the conditions for cognitive mediation. After experimentation with the

[2] Eurocentres: Foundation for European Language and Educational Centres. A Swiss foundation teaching languages worldwide in regions where they are spoken since 1960, an INGO with a participatory status to the Council of Europe since 1968, organiser of the Council of Europe intergovernmental symposia at Rüschlikon in 1971 and 1991, proposer of the European Language Portfolio (ELP): played a key role in the development of the CEFR descriptors and the ELP, see www.eurocentres.com

[3] Brian North, Tim Goodier, Tunde Szabo

[4] Gilles Breton, Hanan Khalifa, Sauli Takala, Christine Tagliante

[5] Coreen Docherty, Gudrun Erikson, Peter Lenz, David Little, Daniela Fasoglio, Neil Jones, Enrica Piccardo, Günther Schneider, Joseph Sheils, Barbara Spinelli, Bertrand Vittecoq

[6] Brian North, Coreen Docherty, Tim Goodier, Hanan Khalifa, Ángeles Ortega, Enrica Piccardo, Maria Stathopoulou, Sauli Takala

categories like pedagogic, cultural, social, and linguistic mediation mentioned above, this distinction between cognitive and relational mediation was adopted as the basis for the mediation categories.

There is, however, no close relationship between the text (Coste and Cavalli 2015) and the descriptors, despite an interplay between the two developments in a series of co-ordination meetings. The new illustrative descriptors for mediation, for example, relate to all four domains of language use defined in the CEFR: public, personal and professional as well as the educational domain, the latter being the focus for Coste and Cavalli. Examples for each of these four domains are in fact given in an elaborated version of the descriptors for mediation activities. Secondly, the mediation project has produced descriptors with a language focus that can be calibrated to a particular CEFR proficiency level, whereas Coste and Cavalli are often concerned with knowledge and values that cannot be related to language proficiency. For this reason, they cite in their text more descriptors from the parallel project concerning Competences for a Culture of Democracy (CCD) than from the mediation project. Nevertheless, their text had a strong influence on the descriptor development, providing the fundamental distinction between cognitive mediation (constructing or conveying meaning) and relational mediation (facilitating relationships) and inspiring many of the actual categories. The list that follows gives the categories at the end of the two years of development.

Mediation activities

Relational mediation

- Establishing a positive atmosphere
- Creating pluricultural space
- Facilitating collaborative interaction
- Managing interaction
- Resolving delicate situations and disputes

Cognitive mediation: Constructing meaning

- Collaborating to construct meaning
- Generating conceptual talk

Cognitive mediation: Conveying received meaning (spoken)

- Relaying specific information
- Explaining data (e.g. in graphs, diagrams, charts etc.)
- Processing text
- Interpreting
- Spoken translation of written text (Sight translation)

Cognitive mediation: Conveying received meaning (written)

- Relaying specific information
- Explaining data (e.g. in graphs, diagrams, charts etc.)
- Processing text
- Translating

Mediation strategies

- Linking to previous knowledge
- Amplifying text
- Streamlining text
- Breaking down complicated information
- Visually representing information
- Adjusting language

Other new scales for related categories

Online interaction (CEFR Section 4.4.3, Council of Europe 2001a:73–87)

- Online conversation and discussion
- Goal-oriented online transactions and collaboration

Text (CEFR Section 4.6.3, Council of Europe 2001a:95–97)

- Expressing a personal response to literature and art
- Analysis and criticism of literature and art

Plurilingual and pluricultural competences (CEFR Section 6.1.3, Council of Europe 2001a:133–135)

- Exploiting plurilingual repertoire
- Exploiting pluricultural repertoire

As can be seen the categories in the last section 'Other new scales for related categories' concern aspects that might well not be considered to be mediation, but in which an element of mediation is involved. Users had requested descriptor scales for literature and for online interaction and, in their work on mediation, the group felt that plurilingual and pluricultural competences were clearly relevant to mediation in a cross-linguistic context. The process of developing and validating these descriptors is described by North and Docherty (this issue).

Future developments

The extended set of CEFR illustrative descriptors, including those from the mediation project, will be presented to users in an extended consultative process before revision for publication, following the precedent set with the CEFR itself. The process will start in June 2016 with a consultative meeting involving experts from the related Council of Europe projects concerning further CEFR developments, languages of schooling and plurilingual and intercultural education. This will be followed by a period of wider consultation and piloting until later 2017. The exact form of presentation and publication will be one of the issues considered in the consultation process.

The Language Policy Unit very much hopes that the new set of descriptors will answer the requests from the field to address the gaps in the original illustrative descriptors and to update them to take account of educational and technological developments. The main set of descriptors will be accompanied by a version collating available descriptors for different age groups of younger learners and relating these to the main set. The research undertaken demonstrates that the validity of the original set of illustrative descriptors is confirmed, despite the passage of time. The new set extends but does not replace them. As explained in this article, one of the surprising and impressive characteristics of the 2001 descriptors is the way that the placement of the vast majority of the original descriptors

on the mathematical scale underlying the CEFR levels has remained constant. Related concepts in the descriptors for new and very different categories have also been calibrated in relation to the original descriptors in a way that appears to be completely coherent.

The descriptors for mediation may have considerable relevance to related projects like the LIAM project and those concerning the languages of schooling. Whereas the original CEFR illustrative descriptors were clearly targeted at secondary school and adult learners of foreign languages, the mediation descriptors have, at least potentially, a broader application, particularly in relation to the teaching and learning of languages across the curriculum, including the language of schooling. Indeed, they could well be of interest to researchers who may be interested in examining the feasibility of developing descriptors for languages of schooling at different levels.

References

Alderson, J C A (Ed) (2002) *Case Studies in the Use of the Common European Framework*, Strasbourg: Council of Europe.

Alderson, J C A and Urquhart, A H (1984) *Reading in a Foreign Language*, London: Longman.

Barnes, D and Todd, F (1977) *Communication and Learning in Small Groups*, London: Routledge and Kegan Paul.

Beacco, J-C (2010) *Items for a Description of Linguistic Competence in the Language of Schooling Necessary for Learning/Teaching History (End of Obligatory Education). An Approach with Reference Points*, Strasbourg: Council of Europe.

Beacco, J-C and Byram, M (2007) *Guide for the Development of Language Education Policies in Europe: From Linguistic Diversity to Plurilingual Education*, Strasbourg: Council of Europe.

Beacco, J-C and Porquier, R (2008) *Niveau A2 pour le français: Un référentiel*, Paris: Didier.

Beacco, J-C, Porquier, R and Bouquet, S (2004) *Niveau B2 pour le français: Un référentiel*, Paris: Didier.

Beacco, J-C, De Ferrari, M, Lhote, G and Tagliante, C (2006) *Niveau A1.1 pour le français / référentiel DILF livre*, Paris: Didier.

Beacco, J-C, Porquier, R and Bouquet, S (2007) *Niveau A1 pour le français: Un référentiel*. Paris: Didier.

Beacco, J-C, Fleming, M, Gouiller, F, Thürmann, E and Vollmer, H (2010) *The Language Dimensions in All Subjects: A Handbook for Curriculum Development and Teacher Training*, Strasbourg: Council of Europe.

Beacco, J-C, Coste, D, Linneweber-Lammerskitten, H, Pieper, I, van de Ven, P-H, Vollmer, H J (2015) *The Place of Languages of Schooling in Curricula*, Strasbourg: Council of Europe.

Beacco J-C, Byram, M, Cavalli, M, Coste, D, Egli Cuenat, M, Goullier, F and Panthier, J (2015) *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education*, Strasbourg: Council of Europe.

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

Breen, M P and Candlin, C N (1980) The essentials of a communicative curriculum in language teaching, *Applied Linguistics* 1 (2), 89–112.

Brown, G, Anderson, A, Shillcock, R and Yule, G (1984) *Teaching Talk: Strategies for Production and Assessment*, Cambridge: Cambridge University Press.

Brumfit, C (1984) *Communicative Methodology in Language Teaching. The Roles of Fluency and Accuracy*, Cambridge: Cambridge University Press.

Brumfit, C (1987) Concepts and categories in language teaching methodology, *AILA Review* 4, 25–31.

Byram, M and Parmenter, L (Eds) (2012) *The Common European Framework of Reference: The Globalisation of Language Policy*, Bristol: Multilingual Matters.

Centre international d'études pédagogiques (2008) *Spoken Performances Illustrating the 6 levels of the CEFR*, Paris: Centre international d'études pédagogiques.

Centre international d'études pédagogiques and Eurocentres (2005) *DVD de productions orales illustrant pour le français, les niveaux du Cadre européen commun de référence pour les langues du Conseil de l'Europe*, Paris: Centre international d'études pédagogiques and Eurocentres/ Didier.

Coste, D and Cavalli, M (2015) *Education, Mobility, Otherness: The Mediation Functions of Schools*, Strasbourg: Council of Europe.

Council of Europe (1992) *Transparency and Coherence in Language Learning in Europe: Objectives, assessment and certification. Symposium held in Rüschlikon, 10–16 November 1991*, Strasbourg: Council for Cultural Co-operation.

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

Council of Europe (2001b) *Un cadre de référence pour les langues: apprendre, enseigner, évaluer*, Paris: Didier.

Council of Europe (2003) *Relating Language Examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment (CEFR)*, Strasbourg: Council of Europe.

Council of Europe (2007) *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities*, paper presented at Intergovernmental Language Policy Forum, Strasbourg, 6–8 February 2007.

Council of Europe (2009a) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*, Strasbourg: Council of Europe, available online: www.coe.int/lang-CEFR

Council of Europe (2009b) *Reference Supplement to the Manual for Relating Language Examinations to the CEFR*, Strasbourg, Council of Europe.

Council of Europe (2011) *Manual for Language Test Development and Examining For use with the CEFR*, available online: www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

Eurocentres and Migros Club Schools (2004) *CEF Performance Samples for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. English. Swiss Adult Learners*, Council of Europe/Migros Club Schools.

Evaluation and Accreditation of Quality in Language Services and Centre international d'études pédagogiques (2015) *Inventaire linguistique des contenus clés aux niveau du CECR*, Paris: Centre international d'études pédagogiques.

Færch, C and Kasper, G K (Eds) (1983) *Strategies in Interlanguage Communication*, Harlow: Longman.

Figueras, N and Takala, S (2016) *Reading and Listening Test Tasks and Items Related to the CEFR*, Strasbourg: Council of Europe.

Glaboniat, M, Müller, M, Rusch, P, Schmitz, H and Wertenschlag, L (2005) *Profile deutsch A1 – C2. Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel*, München: Langenscheidt.

Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*, English

Profile Studies volume 2, Cambridge: UCLES/Cambridge University Press.

Instituto Cervantes (2007) *Niveles de Referencia para el español, Plan Curricular del Instituto Cervantes*, Madrid: Biblioteca Nueva.

Lado, R (1961) *Language Testing*, London: Longman.

Linneweber-Lammerskitten, H (2010) *Items for a Description of Linguistic Competence in the Language of Schooling Necessary for Learning/Teaching Mathematics (End of Obligatory Education). An Approach with Reference Points*, Strasbourg: Council of Europe.

Lüdi, G (2014) Dynamics and management of linguistic diversity in companies and institutes of higher education: Results from the DYLAN project, in Gromes, P and Hu, A (Eds) *Plurilingual Education: Policies – practices – language development*, Hamburg Studies on Linguistic Diversity 3, Amsterdam: John Benjamins, 113–138.

Martyniuk, W (Ed) (2010) *Relating Language Examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual*, Studies in Language Testing volume 33, Cambridge: Cambridge University Press.

Mercer, N and Hodgkinson, S (Eds) (2008) *Exploring Talk in Schools*, London: Sage.

North, B (1991) Standardisation of continuous assessment grades, in Alderson, J C and North, B (Eds) *Language Testing in the 1990s*, London: Macmillan/British Council, 167–177.

North, B (1992) European Language Portfolio: Some options for a working approach to design scales for proficiency, in Schärer, R and North, B (Eds) *Towards a Common European Framework for Reporting Language Competency*, Washington, DC: NFLC Occasional Paper, National Foreign Language Center, 158–172.

North, B (1993a) *The Development of Descriptors on Scales of Proficiency: Perspectives, Problems, and a Possible Methodology*, Washington, DC: NFLC Occasional Paper, National Foreign Language Center.

North, B (1993b) L'évaluation collective dans les Eurocentres, *Le Français dans le Monde – Récherches et Applications*, 69–81.

North, B (1995) The development of a common framework scale of descriptors of language proficiency based on a theory of measurement, *System* 23 (4), 445–465.

North, B (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang.

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.

North, B and Schneider, G (1998) Scaling descriptors for language proficiency scales, *Language Testing* 15 (2), 217–262.

North, B, Page, B, Porcher, L, Schneider, G and Van Ek, J A (1992) *A Preliminary Investigation of the Possible Outline Structure of a Common European Language Framework*, Strasbourg: Council of Europe.

North, B, Ortega, Á and Sheehan, S (2010) *British Council – EAQUALS Core Inventory for General English*, London: British Council/Eaquals.

Oscarson, M (1979) *Approaches to Self-assessment in Foreign Language Learning*, Oxford: Pergamon.

Oscarson, M (1984) *Self-assessment of Foreign Language Skills: A Survey of Research and Development Work*, Strasbourg: Council of Europe.

Oxford Certificate of Educational Achievement (1984) *Oxford Certificate of Educational Achievement: The Personal Record Component: A Draft Handbook for Schools*, Oxford: Oxford Certificate of Educational Achievement.

Parizzi, F and Spinelli, B (2009) *Profilo della Lingua Italiana*, Firenze: La Nuova Italia.

Pieper, I (2010) *Items for a Description of Linguistic Competence in the Language of Schooling Necessary for Learning/Teaching Literature (End of Obligatory Education). An Approach with Reference Points*, Strasbourg: Council of Europe.

Porto, M (2012) Academic perspectives from Argentina, in Byram, M and Parmenter, L (Eds) *The Common European Framework of Reference: The Globalisation of Language Policy*, Bristol: Multilingual Matters, 129–138.

Schärer, R and North, B (Eds) (1992) *Towards a Common European Framework for Reporting Language Competency*, Washington, DC: NFLC Occasional Paper, National Foreign Language Center.

Schneider, G and North, B (2000) *Fremdsprachen können: was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*, Nationales Forschungsprogramm 33, Chur/Zurich: Wirksamkeit unserer Bildungssysteme, Verlag Ruegger.

Schneider, G, North, B and Koch, L (2000) *A European Language Portfolio*, Bern: Berner Lehrmittel- und Medienverlag.

Stathopoulou, M (2015) *Cross-Language Mediation in Foreign Language Teaching and Testing*, Cleveland: Multilingual Matters.

Stern, H H (1983) *Fundamental Concepts of Language Teaching*, New York: Oxford University Press.

Swales, J M (1990) *The Genre Analysis: English in Academic and Research Settings*, Cambridge: Cambridge University Press.

Tarone, E (1983) Some thoughts on the notion of 'communication strategy', in Færch, C and Kasper, G K (Eds) *Strategies in Interlanguage Communication*, Harlow: Longman, 63–68.

Trim, J L M (1978) *Some Possible Lines of Development of an Overall Structure for a European Unit/Credit Scheme for Foreign Language Learning by Adults*, Strasbourg: Council of Europe.

van Ek, J A (1976) *The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults*, Strasbourg: Council of Europe.

van Ek, J A (1986) *Objectives for Foreign Language Teaching. Volume I: Scope*, Strasbourg: Council of Europe.

van Ek, J A and Trim, J L M (2001a) *Waystage*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J L M (2001b) *Threshold 1990*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J L M (2001c) *Vantage*, Cambridge: Cambridge University Press.

Vollmer, H J (2010) *Items for a Description of Linguistic Competence in the Language of Schooling Necessary for Learning/Teaching Sciences (End of Obligatory Education). An Approach with Reference Points*, Strasbourg: Council of Europe.

Wall, J A and Dunne, T C (2012) Mediation research: A current review, *Negotiation Journal* 28 (2), 217–244.

Wilds, C P (1975) The oral interview test, in Spolsky, B and Jones, R (Eds) *Testing Language Proficiency*, Washington DC, Center for Applied Linguistics, 29–44.

Wilkins, D A (1978) Proposal for levels definition, in Trim, J L M (1978) *Some Possible Lines of Development of an Overall Structure for a European Unit/Credit Scheme for Foreign Language Learning by Adults*, Strasbourg: Council of Europe, 71–78.

Zarate, G A, Gohard-Radenkovic, A and Lussier, D (2004) *Cultural Mediation in Language Learning and Teaching*, Graz: European Centre for Modern Languages.

Validating a set of CEFR illustrative descriptors for mediation

BRIAN NORTH EUROCENTRES FOUNDATION, SWITZERLAND
COREEN DOCHERTY RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

This article outlines in more detail the procedures followed and the results achieved in the second part of the project to update the Common European Framework of Reference (CEFR, Council of Europe 2001) illustrative descriptors referred to by North and Panthier (this issue) as 'Updating CEFR descriptors: Phase 2: 2014–2016 – Focus on Mediation'. Mediation was introduced into the CEFR as the fourth mode of communicative language activity by Daniel Coste, further developing North's (1992) fourth category 'processing'. Mediation is the most complex of the four (reception, interaction, production, mediation) because it usually encompasses all of the other three, together with a cognitive and interpersonal challenge. This is because mediation is not concerned with one's own needs, ideas or expression, but with those of the party or parties for whom one is mediating. Mediation thus tends to involve a self-effacing bridging effort to 'get something across' and facilitate the (mutual) understanding of other people. Coste and Cavalli (2015:12) take this idea a stage further and claim: 'In all cases, the aim of the mediation process, defined in the most general terms, is to reduce the gap between two poles that are distant from or in tension with each other.' However, in its treatment of mediation, the CEFR maintained the original focus on 'processing' in the sense of summarising and/or explaining to another person the content of a text to which they do not have access, often because of linguistic, cultural, semantic or technical barriers. An emphasis on mediation as cross-linguistic information transfer is maintained in the many descriptors produced for Profile Deutsch (Glaboniat, Müller, Rusch, Schmitz and Wertenschlag 2005), in innovative cross-linguistic test tasks developed in Greece (Stathopoulou 2013), and in a new, optional plurilingual Matura oral examination offered in Austria (Piribauer, Steinhuber, Atzlesberger, Mittendorfer, Ladstätter, Greinix and Renner 2014).

However, unlike with reception, interaction and production, no descriptors for mediation had been developed and validated in a way comparable to the CEFR illustrative descriptors. Therefore descriptors needed to be developed, validated and calibrated from scratch, so the Mediation Working Group[1] was set up to do so. Since descriptors were to be developed with no precedent, it was decided to adopt a broader interpretation of the concept in line with developments in the wider educational field. In this respect, the reflections of Piccardo (2012) on the way in which the CEFR had anticipated but not developed the concept was very helpful. The resultant set of categories, given in North and Panthier (this issue), was developed over time. The set includes the construction, as opposed to just the conveyance, of meaning, relational mediation (i.e. mediating to manage interpersonal relationships) as well as cognitive mediation, and mediation strategies as well as mediation activities. With regard to the strategies, the work of Stathopoulou (2015) was useful as a starting point.

Descriptor development

The approach taken to the development and validation of the descriptors was based on the one adopted for the development of the original CEFR illustrative descriptors (Council of Europe 2001:218, North and Schneider 1998). This followed a three-phase cumulative process:

- intuitive phase: collecting and reviewing relevant source material, editing existing descriptors and drafting new ones following the guidelines outlined in the CEFR (positive, brief, clear, independent, definite), sifting, classifying, discussing and editing in an iterative process
- qualitative phase: workshops with teachers evaluating and judging descriptors and matching them to the category they were intended to describe
- quantitative phase: calibration of the best descriptors on the basis of a Rasch analysis of the data from the descriptors being used for (self-) assessment, the descriptors being presented in a set of overlapping questionnaires.

There were, however, several differences between the work being reported and the original research (North 2000, North and Schneider 1998):

- Intuitive phase: In the current project, any mediation descriptors from CEFR-related projects took an information-transfer (i.e. interpretation, translation) view of mediation, so the vast majority were inspired by wider reading, rather than coming from existing scales.
- Qualitative phase: Whereas in the original CEFR-descriptor research, all 32 workshops of 2–20 teachers had to be face-to-face, the existence of the internet – plus the networks of organisations like European Association for Language Testing and Assessment (EALTA), Evaluation and Accreditation of Quality in Language Services (Eaquals), European Confederation of Language Centres in Higher Education (CERCLES) and UNIcert (a German organisation focused on university language education and certification) – meant that 137 workshops could be carried out at a distance, with materials emailed to co-ordinators. This Phase 1 of the validation process is described in detail later in this article.
- Quantitative phase: Whereas CEFR levels and descriptors did not exist in the original CEFR-descriptor research, in

[1] Brian North (project co-ordinator), Coreen Docherty, Tim Goodier, Hanan Khalifa, Ángeles Ortega, Enrica Piccardo, Maria Stathopoulou and Sauli Takala.

the current project the networks mentioned above were all familiar with them. Therefore, in addition to the original rating task (using a 0–4 rating scale to answer the question 'Could the person concerned, do what is described in the descriptor?'), it was possible to ask informants to match descriptors to levels, as in a standard-setting task for test items or a standardisation session with video samples. This allowed the use of two complementary ways of calibrating descriptors, in Phases 3 and 2 respectively, as described later in this article.

First, however, a collection of descriptors was necessary as a starting point. The first author (Brian North) first put together descriptors from a range of articles and existing scales, grouping these under provisional headings. This collection was presented to experts at a consultation meeting held in Strasbourg in June 2014 to get informed feedback on the work undertaken up to this point. As a result of this feedback, the collection was then revised between July and September 2014 in preparation for a first meeting of the Mediation Working Group at the end of September. The group then reviewed the descriptors in an iterative manner in a series of workshops and email exchanges held between then and February 2015. The result was an initial set of 427 descriptors for validation, organised into 24 categories. As mentioned by North and Panthier (this issue), six of those categories concerned the following three areas peripheral to the mediation concept: online interaction, responding to literature and art, and plurilingual and pluricultural competences.

Scale and descriptor validation

Once the draft descriptors were finalised, a series of validation activities were undertaken in order to ensure that the individual scales are coherent, and that related scales can be differentiated from each other, before calibrating the descriptors following the Rasch methodology adopted for the original CEFR illustrative descriptors (North 2000, North and Schneider 1998). The validation process was organised in three phases:

- Phase 1: Allocating descriptors to categories
- Phase 2: Assigning descriptors to CEFR levels
- Phase 3: Rating a person's ability to perform what is described by a descriptor.

The validation activities were carried out between February and November 2015 and the data collected from each phase informed decisions to revise, delete and further develop the descriptors as necessary.

Phase 1: Allocating descriptors to categories

The first phase of validation focused on determining the extent to which scales represented distinct categories. For example, could participants distinguish descriptors from scales which may seem related such as 'collaborating to construct meaning' (cognitive mediation) and 'facilitating collaborative interaction with peers' (relational mediation), which both focus on collaboration but from different angles. In addition, the quality of the draft descriptors was evaluated in terms of clarity, usefulness and authenticity. A call to participate was sent out to key contacts in institutes around the world involved in language education and assessment. Those who indicated an interest were asked to organise a 3-hour workshop with interested colleagues and to distribute information about the project to them. The main task during the workshop was for pairs of participants to identify the intended category of descriptors, to rate them for clarity, pedagogical usefulness and relation to real-world language use, and to suggest improvements to the wording.

A total of 472 descriptors were distributed in a series of 30 overlapping sets, with sets being allocated to different institutes. Each scale was included in at least three sets, allowing for the scale descriptors to be considered alongside six or seven other scales in total. Certain categories of existing CEFR descriptors on related areas were also included. Pairs of participants were asked to discuss and rate one of the 30 different sets of about 60 descriptors, presented in random order. An example of the data collection worksheets used by participants can be seen in Figures 1 and 2. In this case, the set focused on the two new online scales. Figure 1 shows

Figure 1: Data collection worksheet

Step 1: Tick in **one** column only (i.e. for a category, or for "Can't Decide" or "Drop this descriptor"), referring to the Descriptor Sheet. Pay attention to the item numbers!!
If you tick "Drop this descriptor," proceed to the next descriptor. Do not do Step 2 in this case. If you tick "Can't decide" you may still do Step 2 if you wish.
Step 2: Judge the quality of the descriptor. Tick Y/N (=**Yes or No)** in the column each of the 3 criteria accordingly.
Step 3: (Optional): If you want to suggest changes to a descriptor, write these directly on the Descriptor Sheet, put your name on the sheet and return it to the coordinator.

Item no.	Step 1: Category (please tick ONE only)				Step 2: Quality (please tick)					
	Online Conversation and Discussion	Goal-oriented Online Transactions and Collaboration	Can't decide	Drop this descriptor	Clearly formulated		Pedagogically useful		Relevant to real life	
					Y	N	Y	N	Y	N
114	✓				✓		✓		✓	
140			✓		✓		✓		✓	

Figure 2: Descriptor worksheet

Item no.	Descriptor
175	Can critically evaluate online comments, embedded links and media, and express negative reactions diplomatically.
177	Can deal ~~confidently effectively~~ with linguistic and cultural problems ~~or cultural issu, es adjusting his/her register appropriately.~~ ~~that arise~~ in order to complete online collaborative or transactional exchanges, ~~adjusting his/her register appropriately.~~

the worksheet used for allocating a descriptor to a scale and evaluating its quality, and Figure 2 shows the worksheet for suggesting reformulation.

Very many informants did suggest reformulations, often striking through or radically editing subordinate clauses. The Group had found it a challenge to get descriptors for mediation down to the 20–25-word length North (2000:345) had discovered teachers had a preference for, and this feedback was as a result invaluable in achieving that aim.

A report was created which collated, for each descriptor, the responses from each set on which the descriptor had appeared. Mediation strategies had appeared on Sets 11, 16, 17 and 20, so in the example in Table 1 for Descriptor 230, the entries for those sets are shown one after another. This descriptor was from the scale for *Linking to previous knowledge* (LINK) and, as can be seen from Table 1, it was overwhelmingly allocated to the correct category.

In order to evaluate the data, coefficients (as percentages) were calculated, following Eichelmann (2015). Table 2 shows the coefficients for assignment to the correct category (OKCoeff), for dropping the descriptor (DRCoeff) and for the three quality coefficients (Clarity, Pedagogical usefulness, Relation to real world). The drop coefficient and the three quality coefficients were also aggregated into an overall coefficient again expressed as a percentage, and shown in larger print, as in Table 2. A subjective criterion was established for each coefficient. For the OK coefficient 50% was adopted, again following Eichelmann (2015), and for the three quality coefficients a higher criterion of 70%. After some thought, 15% was adopted for the drop coefficient. These criteria worked well for distinguishing possible problems. In Table 2, the unsatisfactory results are in bold. From the data in Table 2, we can deduce that although Descriptor 230 was well allocated to its category (see Table 1) it is not overly popular. It hits the drop criterion on two of the four sets, but in the aggregate result (percentage of total pairs) ends just below the 15%. It just fails the criterion of pedagogical usefulness, in addition to that for clarity. After discussion, we dropped it.

Approximately 990 participants (or 495 pairs) from 137 institutes from around the world took part in the Phase 1 workshops. The participating institutes were organised into five 'divisions' of approximately 30 groups each, in order to ensure some representativeness. One division was based on Eaquals (Evaluation and Accreditation of Quality in Language Services), one on members of CERCLES (European Confederation of Language Centres in Higher Education), one on German and American institutes (expected to be less familiar with the CEFR), and the other two were of mixed nationality.

Although all phases of the project required relatively intensive co-ordination in order to maximise participation, the processing of data collected in this phase was particularly labour intensive as co-ordinators sent in electronic copies or photographs of the data collection worksheets, which needed to be manually entered into an Excel spreadsheet. In addition, an unexpected factor which significantly increased the time needed to enter all the data was that somewhere in the region of 200 informants had been anticipated, not 900! A major improvement in the data processing between Phase 1 and the subsequent phases was the use of electronic data collection methods, which eliminated the manual data transfer step.

Phase 2: Assigning descriptors to CEFR levels

The original intention of Phase 2 had been to sort the descriptors into levels, following on from the Phase 1 task of sorting descriptors into categories. Calibration would then follow in Phase 3 with a large online survey. However, the great and growing interest in the project meant that for Phase 2 there would be well over 1,000 respondents allowing for a Rasch analysis to be done. Therefore, it was decided to do a first round of calibration in this phase.

Phase 2 was also conducted as a workshop with participants asked to judge the CEFR level of descriptors. Just over 400 descriptors were presented in a series of 23 overlapping questionnaires created using an online survey tool. All 23 questionnaires started with a common set of 10 CEFR descriptors, which were the main anchor items as these descriptors were calibrated. In addition another nine calibrated

Table 1: Collated data on categories

ID	Set	Scale	Scale comparisons*									Can't decide	Drop it
			GEN	STIM	PROsp	LINK	RESTR	AMPL	STREA	ADJU	INFO		
230	11	LINK				20							4
230	16	LINK		1		9						2	1
230	17	LINK				13		1				2	2
230	20	LINK				12					1	1	3

*GEN = Generating conceptual talk; STIM = Stimulating interaction in plenary and groups; PROsp = Processing text in speech; RESTR = Restructuring text; AMPL = Amplifying text; STREA = Streamlining text; ADJU = Adjusting language; INFO = Information exchange

Table 2: Descriptor coefficients

Serial	OKCoeff	DRCoeff	CLEAR	Coeff	PED	Coeff	REAL	Coeff	Pairs
230	83	**17**	13	**54**	11	**46**	12	**50**	24
230	69	8	10	77	12	92	11	85	13
230	72	11	10	**56**	13	72	15	83	18
230	71	**18**	14	82	14	82	14	82	17
		14		**65**		**69**		72	72

CEFR descriptors for 'Cooperating' and for 'Sociolinguistic appropriateness' were included on certain scales, together with five CEFR descriptors for 'Processing'. Prior to completing the survey, participants were given two familiarisation tasks undertaken in pairs. The first familiarisation task involved identifying the levels of the entries in a jumbled, simplified version of CEFR Section 3.6 (Council of Europe 2001:33–36) on the salient features of spoken language at the CEFR levels; the second familiarisation involved identifying the level of 10 CEFR descriptors.

In the survey task, participants entered their decisions on level first on a paper printout of the questionnaire individually and then, after an opportunity for reflection and review, they entered their responses into the online survey. The survey question was: 'At what CEFR level do you think a person can do what is defined in the descriptor?' Participants were given 10 proficiency bands to choose from, emulating the bands created in the original research that created the scale behind the CEFR levels (North 2000, North and Schneider 1998, Schneider and North 2000): pre-A1, A1, A2, A2+, B1, B1+, B2, B2+, C1 and C2.

The decision to offer the 10 proficiency bands, including plus levels, rather than the six criterion levels was taken after much discussion and with some trepidation. Raters are known to be challenged when faced with a rating scale of more than five or six categories; cognitive overload can result in inconsistent ratings. However, the assumption was made that participants were familiar with the CEFR levels, so this was not just any rating scale. Experience in the video benchmarking seminars held by Centre international d'études pédagogiques (CIEP) in Sèvres in 2005 (for French) and in 2008 (cross-linguistic: five languages) suggested that once people are familiar with 10 levels, they have little difficulty distinguishing between them – though they will do so with differing degrees of severity, despite standardisation training.

The 10-band variant was adopted because Levels B1+ and B2+ had seemed particularly real during the process of developing descriptors for mediation, and because one of the aims in producing the Extended Set of Illustrative Descriptors was to more fully flesh out the plus levels, so it seemed best to ask participants to consider them consciously. Descriptors for both criterion and plus levels were included in each of the two familiarisation tasks and in the main anchor items placed at the beginning of each questionnaire.

A total of 189 institutions from 45 countries and 1,294 persons took part in Phase 2. This was fairly remarkable considering that the survey was distributed in May and June, which is an extremely busy time of year for educational and examination institutes. The aim was for each survey to be rated by 40–50 persons so that (given the overlapping sets) each descriptor would be rated by 100 persons. This goal was met for all descriptor scales: the lowest number of respondents for any one scale being 151 and the highest 273.

In addition to the descriptor ratings, information was collected on a range of demographic variables to give the possibility of future Differential Item Functioning (DIF) analysis, which is used to investigate whether a descriptor is interpreted in a significantly different way by different groups of users. The demographic data also allowed greater insight into the profile of participants. For example, the countries with the most respondents were Spain (10%), Germany (10%), Italy (9%) and the UK (8%). Not surprisingly, English was the most common first language (22%) followed by Spanish (12%) and German (10%). As the majority of respondents (78%) had a first language other than English, speakers of romance languages accounted for the largest proportion of respondents (25%) followed by speakers of Slavic languages (16%) and then Germanic languages (13% – this figure does not include English speakers). Of the participants who completed the background questionnaire, most reported they were teachers (52%) with assessment being reported as the next most common profession (17%). The majority of respondents (62%) worked in higher education.

Nearly 70% of respondents indicated that they were very familiar with the CEFR (selecting 7, 8 or 9 on a scale from 0: 'not at all familiar' to 9: 'very familiar'); however, when looking at respondents' level of experience with each CEFR level, we found that they were most familiar with B1 and B2 followed by A2 and then C1. Respondents had the least amount of experience with A1 and C2.

As this task required participants to have some familiarity with the CEFR in order to judge the level, respondents who reported being less familiar with the CEFR (selecting 0–4 on the 10-point scale) and who had no or limited experience with all CEFR levels were removed from the dataset for analysis purposes (2%). Also, because the descriptors were presented in English, anyone who described their level of English as below B2 (another 2%) was also removed from the dataset. When applying these criteria only 46 respondents were removed from the dataset.

Two complementary analysis methods were adopted for the Phase 2 data: (a) collation of raw ratings to percentages, and (b) Rasch analysis (Linacre 2015). For the simple collation, 50% of respondents choosing the same level, without a wide spread across other levels, was taken as a definitive result. Three different approaches to anchoring the Rasch logit scale to the scale underlying the CEFR levels were applied. The first method was to anchor the steps of the 10-level rating scale to the cut-points between CEFR levels on the logit scale reported by North (2000). Although this is the most obvious method and resolved the issue of the results coming with plus and minus reversed, it systematically underestimated the difficulty of the items that had been included as anchors, producing results that paradoxically disagreed with those from the simple collation of responses to percentages. The second method was to anchor the items included from the CEFR to their logit values (North 2000), which appeared to give sensible results. The third method was to leave the analysis unanchored and then subsequently to equate the new scale to the North (2000) scale through a technique based on the difficulty values and standard deviations of the anchor items.[2] This approach gave results very similar to when the anchor items were anchored directly. A fourth logit value was calculated for each descriptor as

[2] The team is very grateful to Michael Corrigan at Cambridge English Language Assessment for timely and effective support on anchoring, including the provision of the precise calculations for the third, equating, method.

the average of the three methods. Where all methods, or all except the first method, agreed on a CEFR level, this was taken as definitive. On that basis, almost 100 of the around 450 descriptors to be calibrated had an identical 'definitive' result from both the collation/percentage and the Rasch analysis. After discussion, the number of descriptors considered calibrated rose to 192.

As with Phase 1, the findings from Phase 2 were used to delete or revise descriptors. In addition, feedback received from participants resulted in a number of new descriptors for A1 and A2 being developed, particularly for 'Exploiting plurilingual and pluricultural repertoires', since it was felt that mediation was very relevant at lower levels of proficiency.

Phase 3: Rating a person's ability

The third validation activity was a large-scale online survey that took place between the beginning of September and the beginning of November 2015. A total of 365 descriptors (including 74 already calibrated anchor items) were once again presented on a series of 23 overlapping questionnaires, with questionnaires being allocated to different institutes. The majority of the anchor items were items that had been calibrated in Phase 2. Each category now had its own anchor items, meaning that it could, if necessary, be analysed separately. In this phase, the questionnaires were created in both English and French to widen participation. In addition to distributing questionnaires to Phase 1 and 2 institutions and to the network of CIEP, links to the questionnaires were made available in an open call through organisations such as the International Federation of Language Teacher Associations (FIPLV), the European Centre for Modern Languages (ECML), Eaquals, EALTA and the Canadian Association of Second Language Teachers (CASLT/ACPLS). Cambridge English and NILE also distributed the open call in English to their networks. Participants were asked to complete the survey more than once, not only about themselves but for different people whom they know well and/or for all the languages they can speak, which gave a total of 3,503 usable responses, with 25% of these responses coming from the French survey. Over 80 countries and 60 languages were represented in the data.

The aim of the survey task was to replicate the original 1994 CEFR scaling activity by asking participants to think about a person that they knew very well (this could be themselves or someone else), and how they perform in a second/foreign language, and to enter a rating against each descriptor. The survey question was as follows: Could you, or the person concerned, do what is described in the descriptor? The same 5-point rating scale that had been used to calibrate the original CEFR illustrative descriptors was also used. The abbreviated form of the scale is given below:

0 Beyond my/his/her capabilities

1 Yes, under favourable circumstances

2 Yes, in normal circumstances

3 Yes, even in difficult circumstances

4 Clearly better than this

Once again, Phase 3 worked very well. The result of the global analysis seemed surprisingly consistent. The only disappointing point was that lower level items for 'Exploiting plurilingual repertoire' tended to come out as B2+. This suggested that respondents were resorting to what is called 'halo effect', giving the same response to a whole series of items without differentiating between them. Apart from these implausible calibrations, the only apparently suspicious-looking ones were for the category 'Expressing a personal response to literature and art'. Some of them seemed to be being placed systematically at levels that appeared to be one level above what one might intuitively expect. An impression of this type could reveal a dimensionality problem. Including data in a single Rasch analysis presupposes technical unidimensionality. This is not at all the same thing as psychological unidimensionality; the Rasch model is very robust and accepts a considerable degree of psychological multidimensionality whilst giving a sensible result. However, where there is a suspicion of a dimensionality problem, categories should be analysed separately to see if this yields different difficulty values (Bejar 1980). This had happened with reading in the original CEFR-descriptor research. Separate analyses were therefore undertaken for all the areas less connected to the central mediation construct:

- For plurilingual and pluricultural competences:
  - creating pluricultural space
  - exploiting pluricultural repertoire
  - exploiting plurilingual repertoire.
- For interpretation and translation:
  - interpreting
  - spoken translation of written text (Sight translation)
  - translation.
- For online interaction:
  - online conversation and discussion
  - goal-oriented online transactions and collaboration.
- For literature and art:
  - expressing a personal response to literature and art
  - analysis and criticism of literature and art.

The separate analyses resulted in some slight changes to calibrations that appeared intuitively sensible, and were closer to the results intended and to those achieved in Phase 2. However, the problem with plurilingual competence remained. It became clear that further work was needed since the bottom half of the scale had not functioned as expected. As a result, further consultation with experts in this area was undertaken and an additional survey focusing on the scales for plurilingual and pluricultural competences is underway.

Results

In total, just over 40% of the complete body of descriptors subjected to the rigorous 3-phase validation process (after initial selection and editing) were, for one reason or another, rejected. However, this included some 30 calibrated descriptors which were removed in the final review in order to reduce repetition, not because of concerns on

quality. The final set presented 367 validated descriptors (before the follow-up on plurilingual and pluricultural competences), calibrated to the scale underlying the CEFR descriptors. There was great consistency in the way that concepts had been scaled to CEFR levels, such as in the following four descriptors on the scale 'Managing interaction', which all concern giving instructions and were calibrated at B2:

> Can explain ground rules for collaborative discussion in small groups that involves problem-solving or the evaluation of alternative proposals.
>
> Can explain the different roles of participants in the collaborative process.
>
> Can give clear instructions to organise pair and small group work and conclude them with summary reports in plenary.
>
> Can intervene when necessary to set a group back on task with new instructions or to encourage more even participation.

Often, the similarity between what was described by descriptors calibrated to the same level required a bit of reflection, as with the following two descriptors also calibrated to B2 with virtually identical logit values in both Phases 2 and 3, for 'Creating pluricultural space' and for 'Exploiting plurilingual repertoire' respectively:

> Can work collaboratively with people who have different cultural orientations, discussing similarities and differences in views and perspectives.
>
> Can alternate between languages in collaborative interaction in order to clarify the nature of a task, the main steps, the decisions to be taken, the outcomes expected.

The fact that it proved possible to calibrate the new descriptors to the scale from the original research (North 2000) was also a considerable achievement. After all, the areas being described were very different (mediation rather than interaction/production), the type of informants was substantially different (mainly university teachers rather than secondary school teachers), they came from 45 countries rather than just from Switzerland, and finally the not inconsiderable fact that the survey took place 20 years later. However, the process of relating the new scale to the original 1994/95 scale was not straightforward and certainly not automatic. Judgement was required. In Phase 2 (assigning descriptors to levels), as described above, three different methods of anchoring to the 1994/95 scale were tried, the first of which gave significantly different difficulty values on the logit scale. Then, for very many items, we had independent logit values from both Phase 2 and Phase 3, which might also differ. The approach taken was to review and take into account all the evidence for each descriptor, as shown in Table 3 for the two plurilingual/pluricultural items mentioned above.

Table 3: Item records

Descriptor	PLUC17collab	PLUL01task-r
Level	B1	B2
Phase 3 Global	B2	B2
Logit value	1.39	1.26
Phase 3 Separate	B2	B2
Logit value	1.40	1.31
Phase 2 Collation	B2	
Phase 2 Rasch	B2/B2+	
Decision	B2	B2

The first descriptor (collaborating with people from different cultures) had originally been intended to be a B1 descriptor, but the two independent placements at B2, in Phase 2 and 3 respectively, were persuasive.

The extent of difference in the interpretation of the difficulty of the descriptors in relation to different demographic variables in this project was comparable to that found in the original research, with around 14% of the descriptors being affected. However, whereas in the original CEFR-descriptor research there had been no statistically significant differences of interpretation by teachers of English, French and German, this time almost all the DIF that came to light did so in the comparison of the judgements of teachers of just French with those of teachers who taught *several* languages, or who just taught English. It may be that the fact that the French version was only available for the Phase 3 open call had a strong influence: the majority of the other informants had been through a series of workshops with the descriptors in Phases 1 and 2.

Conclusion

The project shows that informants can interpret consistently the difficulty level of descriptors for mediation, including descriptors for aspects of pluricultural competence; plurilingual competence, a more recent concept, is more problematic. One expert on plurilingualism (Barbara Spinelli) had in fact warned that there would be problems trying to calibrate descriptors with informants who were not working on plurilingualism and certainly the halo effect in the responses suggests cognitive overload. It probably did not help that these descriptors came right at the end of a long survey too. The follow-up will be undertaken with two groups: known experts in plurilingualism and volunteers without knowledge of plurilingual issues, in an attempt to address this issue.

The response to the project also shows that there is a considerable enthusiasm for further development and research related to the CEFR. The Council of Europe's name obviously helped, but it is remarkable that approaching 1,000 people took part in all three validation phases. The very diverse groups of respondents clearly valued the opportunity to participate. A total of nearly 1,000 comments were made by participants in Phase 2 (631) and Phase 3 (364), many of which were comprehensive and insightful, indicating a high level of engagement with the task. After Phase 2, 93% of the informants had stated they would be interested to continue and even after Phase 3, 76% indicated that they would like to participate in similar future projects.

Above all, what the project showed was that in the development of descriptors it really is indispensable to undertake qualitative validation, as in the original research (North 2000) and in some recent projects (e.g. Eichelmann 2015, Vogt 2011). In such validation, groups of informants who are independent of the development team and representative of the end users sort, evaluate and suggest reformulations of the descriptors, as in our Phase 1. On linking to the CEFR, the experience with the analysis underlines the message of the Council of Europe's 2009 Manual, *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* that complementary standard-setting methods

should be used (as in Phases 2 and 3) and that all the evidence should be taken into account in order to make a considered judgement.

Following the review process at the completion of the analyses, a series of consultations will be undertaken before the descriptors are circulated more widely in a preliminary, consultative edition late in 2016. Following the procedure adopted with the CEFR itself in 1997, such piloting will be accompanied by feedback questionnaires. The process of the development and validation of the descriptors is also thoroughly documented, as with the original set. The exact form of the final presentation of the new descriptors will be one of the questions in the consultation process, but from autumn 2016 they should be available in their preliminary, pilot form on the Council of Europe's website. The development group would like to express their gratitude to all those institutions and individuals who made that possible.

References

Bejar, I (1980) A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates, *Journal of Educational Measurement* 17 (4), 283–296.

Coste, D and Cavalli, M (2015) *Education, Mobility, Otherness: The Mediation Functions of Schools*, Strasbourg: Council of Europe DGII – Directorate General of Democracy, Language Policy Unit.

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

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*, Strasbourg: Council of Europe.

Eichelmann, T (2015) *Der gemeinsame europäische Referenzrahmen für Sprachen und die Hochschulspezifik am Beispiel UNIcert*, unpublished PhD thesis, University of Dresden.

Glaboniat, M, Müller, M, Rusch, P, Schmitz, H and Wertenschlag, L (2005) *Profile deutsch A1 - C2. Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel*, München: Langenscheidt.

Linacre, J M (2015) *Winsteps: Rasch-model Computer Program*, Chicago: MESA Press.

North, B (1992) European Language Portfolio: Some options for a working approach to design scales for proficiency, in Council of Europe *Transparency and Coherence in Language Learning in Europe: Objectives, assessment and certification. Symposium held in Rüschlikon, 10–16 November 1991*, Strasbourg: Council for Cultural Co-operation, 158–174.

North, B (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang.

North, B and Schneider, G (1998) Scaling descriptors for language proficiency scales, *Language Testing* 15(2), 217–262.

Piccardo, E (2012) Médiation et apprentissage des langues: Pourquoi est-il temps de réfléchir à cette notion? *ELA: Études de Linguistique Appliquée* 167, 285–297.

Piribauer, G, Steinhuber, B, Atzlesberger, U, Mittendorfer, F, Ladstätter, T, Greinix, I and Renner, H (2014) *Wegweiser: Mehrsprachigkeitsprüfungen. Linguistic mediation: How to assess plurilingual oral competences within the final Austrian baccalaureate exam*, unpublished manuscript.

Schneider, G and North, B (2000) *Fremdsprachen können: was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*, Nationales Forschungsprogramm 33, Chur/Zurich: Wirksamkeit unserer Bildungssysteme, Verlag Ruegger.

Stathopoulou, M (2013) The linguistic characteristics of KPG written mediation tasks across levels, in Lavidas, N, Alexiou, T and Sougari, A-M (Eds) *Major Trends in Theoretical and Applied Linguistics: Selected Papers from the 20th ISTAL*, London: Versita de Gruyter, 349–366.

Stathopoulou, M (2015) *Cross-Language Mediation in Foreign Language Teaching and Testing*, Cleveland: Multilingual Matters.

Vogt, K (2011) *Fremdsprachliche Kompetenzprofile*, Tübingen: Narr Verlag.

'Learning through languages' conference of the European Centre for Modern Languages (ECML)

WALDEMAR MARTYNIUK JAGIELLONIAN UNIVERSITY, POLAND

On 10–11 December 2015 the European Centre for Modern Languages (ECML) of the Council of Europe in Graz (Austria) presented the results of its 2012–2015 programme 'Learning through languages: Promoting inclusive, plurilingual and intercultural education' and launched its new 2016–2019 programme entitled 'Languages at the heart of learning'.[1] Conference participants – representatives of the 33 ECML member states and other invited partners – were presented with the outputs of the 2012–2015 ECML programme of activities and invited to discuss how to disseminate and implement these resources in different educational contexts.

The publications, websites and applications resulting from the 2012–2015 programme address key issues in contemporary language education such as support for learners from a migrant background, the development of teachers' digital skills or the need to support the development of language competences alongside subject-related competences. A selection of these new resources are briefly presented in the next section.

The conference was also an occasion to celebrate 20 years of extensive work in the field of language education by the ECML. The centre is an enlarged partial agreement[2] of the Council of Europe comprising 33 European states (as of December 2015).

[1] The conference was streamed via the web and the videos of the individual sessions are now available online at: www.ecml.at

[2] An agreement between member states and one or more non-member states

It was set up in 1994 and started its activities a year later in response to a need expressed by the Council of Europe's member states to enhance quality in language education and reinforce respect for the cultural and linguistic diversity in Europe. Its mission is to assist and support its member states in the implementation of innovative approaches and dissemination of good practice in all three areas of language education: learning, teaching, and assessment. The centre's work is based on an extensive international co-operation with large-scale medium-term programmes of projects as the core activity. Every four years the centre launches a Call for Proposals inviting experts in the field of language education to submit project ideas related to priorities identified by the Centre's member states. The activities of the selected projects are supported by the ECML in terms of financing and administration. Project activities include expert and network meetings as well as workshops to which participants from the member states are invited. The outputs of the projects carried out within a given medium-term programme take the form of training kits, guidelines and interactive websites targeted at teachers, teacher trainers, curriculum developers, and decision-makers – all of them freely available in print or for download.[3] Over the last 20 years the ECML has build up a considerable network of and for a broad range of stakeholders including individual teachers, teacher trainers, researchers, and administrators as well as professional bodies and international non-governmental organisations working in language education. In 2010 the ECML invited leading international non-governmental organisations acting in this field (among them ALTE – the Association of Language Testers in Europe) to set up a Professional Network Forum to share know-how and work together on areas of common interest.

First results of the ECML 'Learning through languages' programme (2012–2015)

The ECML 2012–1015 programme was based on a long-term vision developed by the centre and adopted by the ECML member states in 2011, which aimed to reach beyond foreign language education and encompass all languages within the plurilingual repertoire of each individual learner in a lifelong-learning process. This vision reflects the concepts developed by the Language Policy Unit (formerly Division) of the Council of Europe in Strasbourg (France) and published in the *Common European Framework of Reference for Languages* (CEFR), the *Guide for the development of language education policies in Europe*, the *Guide for the development and implementation of curricula for plurilingual and intercultural education* and a range of other related documents and tools, now parts of an extensive set of resources made available through the virtual *Platform of resources and references for plurilingual and intercultural education*.[4] At the core of the Council of Europe's approach to language education is the view that an adequate development of language competences is a condition for unrestricted and fair access to good-quality education which, in turn, constitutes the necessary basis for the individual, social, and professional success of a learner, and in this way contributes to social cohesion, democratic citizenship, intercultural dialogue, and economic progress of our societies. In the rationale published for the 2012–2015 programme in 2011, long before the dramatic increase in migration and mobility in Europe, the attention was rightly given to the challenges related to the management of growing linguistic and cultural diversity calling for approaches to education that view this phenomenon 'not as an obstacle or a problem, but rather as an asset and a potential benefit to society'. Such approaches would mean 'moving away from the teaching and learning of languages as separate, unrelated and thus isolated (school) subjects. The new task for education envisaged by the ECML document is 'to provide coherent support for the lifelong development of transversal, individual strategies in order to deploy available linguistic resources purposefully, thus making efficient use of one's own range of language competences.'[5] Several projects under the 2012–2015 programme addressed these needs in an innovative way. A selection of them is briefly presented in the next section.

Development projects

PlurCur: Plurilingual whole school curricula, a project co-ordinated by Britta Hufeisen (Germany), piloted and assessed the concept of a school policy comprising majority and minority, regional, heritage and neighbouring languages. The policy 'is designed in such a way that languages taught as subjects are not treated in isolation and language and non-language instruction overlap so that all subject teaching is also language teaching.'[6] The project website offers resources which help to clarify, develop and implement the whole-school policy in different contexts by giving examples of activities piloted in project partner schools and discussing success factors for the implementation of the approach.

A project co-ordinated by Oliver Meyer (Germany) – *A pluriliteracies approach to teaching for learning* – aimed at developing content and language integrated learning (CLIL) approaches with a focus on providing support for academic literacy in secondary education 'to help learners to become better meaning-makers, who can draw on content knowledge to communicate successfully across languages, disciplines, and cultures.'[7] The project website offers informative videos explaining the approach, as well as ideas for putting it into practice.

The project *Maledive – Teaching the language of schooling in the context of diversity*, co-ordinated by Eija Aalto (Finland), focused on teacher education for the majority language (e.g. Swedish in Sweden, Polish in Poland) drawing on the linguistic and cultural diversity in the (multilingual) classroom for the benefit of all learners. The project website offers study materials for pre- and in-service teacher education as well as

[3] www.ecml.at/publications
[4] www.coe.int/lang-platform
[5] www.ecml.at/ECML-Programme/Programme2012-2015/tabid/685/Default.aspx
[6] www.ecml.at/plurcur
[7] www.ecml.at/clilandliteracy

ideas for the promotion of teacher collaboration across school subjects and examples of activities to develop plurilingual approaches.[8]

Educomigrant – Collaborative Community Approach to Migrant Education – a project co-ordinated by Andrea Young (France) explored new ways to enhance young migrants' education by developing links between schools, the home and local partners in education. The project website offers a Virtual Open Course on a Moodle platform as well as strategies and materials for trainers in multilingual educational settings.[9]

The project *Language skills for successful subject learning*, initiated by a team representing institutions with membership to the Association of Language Testers in Europe (ALTE) and co-ordinated by Eli Moe (Norway), produced CEFR-linked language competence descriptors for mathematics and history/civics for learners aged 12/13 and 15/16 which were made available in six languages (English, French, Norwegian, Lithuanian, Portuguese, Finnish).[10]

The *ProSign* project addressed the largely neglected area of sign languages by establishing descriptors and approaches to assessment for sign languages in line with the CEFR. The extensive project website offers definitions of CEFR proficiency levels for sign languages together with a proposal for an assessment cycle for language proficiency in sign languages.[11]

Mediation activities

Some of the ECML activities under the 2012–2015 programme were devoted not to new developments but to dissemination (*mediation* in ECML terms) of project results delivered as an output of the previous work programmes. The tools developed earlier within the CARAP/FREPA (Cadre de Référence pour les Approches Plurielles des Langues et des Cultures/Framework of Reference for Pluralistic Approaches to Language and Culture) project (*A Framework for Pluralistic Approaches to Languages and Cultures*) for example, were disseminated through a series of activities in the majority of the ECML member states. Twenty-two specific country pages, in the language(s) of the country, offer a presentation of the Framework, as well as translations, partial or complete, of FREPA descriptors. An online database of teaching materials and a training kit for teachers are available through the main project website.[12]

Training and consultancy

Another example of an effort to promote and disseminate the results of the many projects co-ordinated and supported by the ECML over the last 20 years (over 80 in total) is an initiative that started under the 2012–2015 programme to offer training and consultancy services to interested member states in areas related to the different projects. The ultimate aim is to turn all the ECML projects – once concluded – to a permanent offer to the member states to provide services ranging from general professional consultancy to targeted training workshops run in the countries by individual experts or expert teams involved in the ECML project work. Two training and consultancy areas under the 2012–2015 programme were offered jointly by the ECML and the European Commission: one providing an extremely useful reviewed inventory of information and communication technology tools and open educational resources (as well as training in the use of them)[13], the other offering assistance and training 'in relating language tests and examinations to the CEFR in a valid and equitable way and in exploring relationships between foreign language curricula and the CEFR.'[14]

Under the new 2016–2019 programme the European Commission offered also to co-finance the ECML training and consultancy service in support of multilingual classrooms. This initiative provides training workshops 'to help member states ensure access to quality education for migrant learners which will help bridge the attainment gap between these learners and non-migrant pupils'.[15] The other areas on offer for training and consultancy under the 2016–19 programme, in addition to the ones mentioned above, include:

- using the *European Portfolio for Student Teachers of Languages* (EPOSTL)[16]
- ensuring quality in language and citizenship courses for adult migrants
- setting up and using an electronic European Language Portfolio (ELP)
- implementing content and language integrated learning (CLIL) approaches
- providing quality education in Romani
- plurilingual and intercultural learning through mobility
- using CARAP/FREPA.

'Languages at the heart of learning': The new medium-term programme of ECML activities

During the conference the new ECML programme 2016–2019 entitled 'Languages at the heart of learning' was officially launched. A panel of experts from different ECML member states shared their views on the ways in which the ECML is expected to address national challenges at European level. The panel discussion was followed by

[8] www.ecml.at/maledive
[9] www.ecml.at/community
[10] www.ecml.at/languagedescriptors
[11] www.ecml.at/prosign
[12] www.ecml.at/carap
[13] www.ecml.at/ictinventory
[14] relang.ecml.at
[15] www.ecml.at/TrainingConsultancy/Multilingualclassrooms
[16] www.ecml.at/epostl

keynote speeches offered by senior representatives of the Council of Europe, the European Commission, and the University of Graz who underlined the need for an extensive European co-operation in the area of language education to build a better, more humane and socially cohesive Europe.

The new ECML programme is structured similarly to the previous one and builds upon three key work strands: Development, Training and Consultancy, and Mediation (dissemination). The Development strand encompasses both already well-formulated project ideas and think tanks where concrete project ideas are expected to emerge from expert discussions. Topics of the developmental projects include digital literacy, reference level descriptions for language teachers, sign language instruction, language education for adult migrants, language training for professional purposes, descriptors for languages of schooling, and quality assurance in the use of the CEFR. A large degree of flexibility is ensured through the Training and Consultancy and Mediation strands, whose exact composition will be determined by member states on an annual basis in the course of the programme.

With the new programme the ECML is following its long-term vision to explore the interdependence between quality education and quality language education, and the recognition that language is at the heart of all learning, which means that all teachers, in languages and other subjects, have an important role to play supporting the development of the linguistic and intercultural repertoire of their learners.

The English Profile Programme 10 years on

FIONA BARKER RESEARCH AND THOUGHT LEADERSHIP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

The English Profile Programme is celebrating its tenth anniversary in 2016, following a decade of collaborative data collection, research, publications and events that have taken place under the English Profile name. This article reports on the latest English Profile network seminar that took place in Cambridge in February 2016, which reflected on some of the highlights of English Profile so far and looked forward to more exciting collaborations to come.

English Profile is a worldwide collaborative network of educational professionals who are all interested in knowing what the Common European Framework of Reference (CEFR, Council of Europe 2001) means for English. The English Profile Project (as it was known for the first few years) aimed to describe English at each CEFR level (A1–C2) in terms of the vocabulary, grammar, functions and other elements that language learners can typically use, building on the 'T-series' volumes published from the 1970s onwards (see van Ek and Trim 1991a/1998a, 1991b/1998b, 2001, also www.englishprofile.org/resources/t-series). The founding English Profile partners, led by Cambridge English Language Assessment and Cambridge University Press, combined their research and development capacity to be recognised by the Council of Europe as the official Reference Level Description (RLD) project for English, thus taking on the task of developing RLDs to complement the work of the other (currently 10) projects developing RLDs for other languages (further detail can be found at: www.coe.int/t/dg4/linguistic/DNR_EN.asp#P30_2633).

The original research agenda of English Profile included three research strands, namely Corpus Linguistics, which involves investigating learner production to identify criterial features at the different levels (that is, features which distinguish one level from another); the Pedagogy strand which focuses on curricula and materials, initially at higher CEFR levels B2–C2; and the Assessment strand, which focuses on how language learners' skills develop and are used at different proficiency levels. Starting from this framework, members of the English Profile Network have sought to answer the following questions over the last decade (adapted from UCLES/Cambridge University Press 2011:7-8):

- How do the different kinds of criterial features (lexical semantic, syntactic, discourse etc.) cluster together to define learner profiles in English?
- Which linguistic features realise which language functions across the CEFR levels?
- How does the profile of the learner vary depending on their L1?
- What are the pedagogical implications of L1 effects for the learning, teaching and assessment of English?
- What are the similarities and differences between adult and young learners of English developmentally?
- What is the role of learner and learning strategies?
- How do all of these factors predict patterns of learner output?
- What type of learning model can accommodate the interactions that underpin language learning?

These and other questions have been explored through various research endeavours over the past decade.

Publications and events

From its launch, English Profile projects have been presented at various ELT and related conferences and events (starting at its launch at IATFEL Harrogate in 2006) and it has had a

successful seminar programme of its own, with the sixteenth event taking place in February 2016. Between 2009–2012 English Profile secured European Commission support and funding through the Lifelong Learning Programme, which enabled it to develop a wider community of teachers and researchers, and which involved high-profile events hosted by the project partners outside the UK along with annual events attracting increasing numbers of participants to events in Cambridge.

Alongside events, English Profile members have been busy producing resources, including the English Vocabulary Profile and the English Grammar Profile which are increasingly widely-cited, along with various publications, all of which are accessible from the English Profile website (www.EnglishProfile.org). Articles from the *English Profile Journal* can also be found online. This journal published peer-reviewed research from the English Profile Programme until 2012 when the first English Profile Studies books were published. This book series currently has five volumes (with two volumes planned for publication each year) which report on research directly related to exploring the CEFR (see Hawkins and Filipović 2012, Green 2012, North 2014) and its impact on teachers and learners (see Ćatibušić and Little 2014, Harrison and Barker (Eds) 2015). All of these resources, together with informative booklets and short videos about the CEFR and English Profile are intended to show not only what aspects of English are typically learned at a particular level, but also how these interact with one another and what they mean for the target audience of teachers, curriculum developers, materials writers and test writers, to provide practical assistance when they need to know what is suitable for learning, teaching or testing English in a particular context.

English Profile Seminar 2016

The most recent English Profile Seminar was the sixteenth to take place, held on 5 February 2016 at the Cass Centre in Cambridge, hosted by Cambridge University Press. Over 100 delegates attended this tenth anniversary seminar which included both established and new English Profile members members, encompassing assessment and publishing colleagues, PhD students, academics in applied and theoretical linguistics and other areas, English teachers, teacher trainers and language school representatives, together representing a range of countries such as Malta, Norway, Italy, Pakistan, the US and Poland. The theme for the day was 'Using learner data to understand language learning and progression' and within this broad theme, three sessions were held which focused on: the theory behind language learning and progression; understanding progression through technology; and finally, the talks broadened out to consider general issues in corpora to focus on learning and progression. After a welcoming speech from Ben Knight, Director of Language Research and Consultancy, ELT at Cambridge University Press, Nick Saville (Director of Research and Thought Leadership at Cambridge English) talked about how Learning Oriented Assessment is the missing link between English language learning and measuring progression, and how technology can help language education (see www.cambridgeenglish.org/research-and-validation/fitness-for-purpose/loa/). Next, Ianthi Maria Tsimpli (University of Cambridge) presented on some language properties that are difficult to learn, even at advanced stages of second language development, also suggesting some shared characteristics that may give rise to this situation, all from second language acquisition studies. To conclude the first section on theories behind language learning, Susan Hunston (University of Birmingham) explored novel approaches to measuring complexity and correctness in learner written output.

The second section of the day focused on technology, with Fiona Barker (Cambridge English) and Sarah Grieves (Cambridge University Press) starting off by reporting on the development and use of the Cambridge Learner Corpus (CLC), a unique corpus of exam scripts from English learners, presenting some key moments from its history. They also announced the planned release of a public subset of the CLC (known as OpenCLC) which will be available for educational use in summer 2016. The next two talks featured Cambridge researchers who reported on experiments on automatically rating written and spoken data. Firstly, Ted Briscoe, Helen Yannakoudakis and Ekaterina Kochmar considered to what extent criterial features are discriminative by using criterial features from the English Grammar Profile to improve a rating system and on the other hand, looking at how a visualisation system can help the user to interpret discriminative features. The third talk in this section by Mark Gales and Kate Knill (University of Cambridge) reported on the machine learning of level and progression in spoken English, also considering how automatic rating systems are improving in their use and power.

The final three talks considered specific aspects of language learning in relation to psycholinguistics, phraseological development and certainty in learner language. Philip Durrant (University of Exeter) presented on the complementarity of learner corpus research and psycholinguistics, showing how ideas and methods from both fields can improve the other. Next, Magali Paquot (University of Louvain) reported on several studies that explored phraseological development in learner English, seeking to assess whether word combinations are 'native-like' using statistical measures and using some of the learner corpora available at the Centre for English Corpus Linguistics at Louvain. Finally, Vaclav Brezina (Lancaster University) presented a corpus-informed study – using the Trinity Lancaster Corpus – which investigated the complex pragmatics of expressing certainty and how task and first language background affects this in language learners' spoken performance.

Michael McCarthy (University of Nottingham, Pennsylvania State University, University of Limerick) drew the day to a close, linking together key strands raised by the speakers and followed up in discussion during the event. What was clear is that English Profile is here to stay and that there remains much work to be done to build on the excellent foundations of the first 10 years of research and collaboration, which we hope that you will consider joining us in, whether focusing on research or practical applications.

References

Ćatibušić, B and Little, D (2014) *Immigrant Pupils Learn English: A CEFR-Related Empirical Study of L2 Development*, English Profile Studies volume 3, Cambridge: UCLES/Cambridge University Press.

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

Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*, English Profile Studies volume 2, Cambridge: UCLES/Cambridge University Press.

Harrison, J and Barker, F (Eds) (2015) *English Profile in Practice*, English Profile Studies volume 5, Cambridge: UCLES/Cambridge University Press.

Hawkins, J A and Filipović, L (2012), *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, English Profile Studies volume 1, Cambridge: UCLES/Cambridge University Press.

North, B (2014) *The CEFR in Practice*, English Profile Studies volume 4, Cambridge: UCLES/Cambridge University Press.

UCLES/Cambridge University Press (2011) *English Profile: Introducing the CEFR for English v.1.1*, available online: www.englishprofile.org/images/pdf/theenglishprofilebooklet.pdf

van Ek, J A and Trim, J L M (1991a/1998a) *Waystage 1990*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J L M (1991b/1998b) *Threshold 1990*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J L M (2001) *Vantage*, Cambridge: Cambridge University Press.

Below: Fiona Barker presenting on corpus resources at the English Profile Seminar 2016

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Studies in Language Testing

An indispensable resource for anyone interested in new developments and research in language testing

To find out more about our full list of publications:

www.cambridge.org/elt/silt

www.cambridgeenglish.org/silt

8664238081

To subscribe to *Research Notes* and download previous issues, please visit:
www.cambridgeenglish.org/research-notes

Contents:

Research Notes

For further information visit the website:
www.cambridgeenglish.org

Cambridge English
Language Assessment
1 Hills Road
Cambridge
CB1 2EU
United Kingdom

www.cambridgeenglish.org/helpdesk

THE QUEEN'S AWARDS FOR ENTERPRISE: 2015

ALTE Association of Language Testers in Europe

A DIVISION OF CAMBRIDGE ASSESSMENT

All details are correct at the time of going to print in April 2016