Constructing comparable standards of communicative language competence: the experience of two European projects

Neil Jones[1]

In diesem Beitrag geht es um die Ziele, Ergebnisse und Werte des Sprachunterrichts. Er befasst sich mit den Ergebnissen zweier empirischer Untersuchungen, dem *European Survey on Language Competences* (ESLC), die der Autor in der zweiten Phase selbst leitete, und der kürzlich erschienenen *Study on Comparability of Language Testing in Europe* (SCLTE), für die der Autor ebenfalls verantwortlich war. In dem Beitrag soll versucht werden, die Schritte zur Konzeptualisierung eines kompetenzorientierten Messverfahrens und dessen Umsetzung zu umreißen. Ausgangspunkt ist die Definition von Konstrukten; in den folgenden Schritten wird dann die Entwicklung eines Testdesigns dargestellt und erläutert, wie ein Test konstruiert werden sollte. Der Aufsatz wird abgerundet durch Erläuterungen zum *standard setting* und zur Konstruktion einer Messskala.

1. The first *European Survey on Language Competences*

In June 2012 the first *European Survey on Language Competences* (ESLC, European Commission 2012a) published its findings, bringing to an end a complex four-year project delivered by a multinational consortium of partners and administered with the assistance of national research coordinators in 16 jurisdictions. The project's sponsor was the European Commission, and Cambridge English Language Assessment was the contracting partner with the Commission.

In calling for the Survey, the Commission's intention was "not only to undertake a survey of language competences but a survey that should be able to provide information about language learning, teaching methods and curricula" (European Commission 2007: 1). The Commission hoped to use the Survey to monitor progress against the March 2002 Barcelona European Council conclusions, which had 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 (European Commission 2005).

Not all countries participated in the *European Survey* (among those who did not was Germany), and one important skill – Speaking – was not included in the Survey because it was judged logistically too difficult by the Commission. None the less, the Survey significantly increased our understanding of the state of lan-

1 Korrespondenzaddresse: Dr. Neil Jones, Cambridge University, neilfjones@ntlworld.com

guage learning at the end of lower secondary education, using the *Common European Framework of Reference* (CEFR, Council of Europe 2001) to report on levels of reading, writing and listening competence, for English, French, German, Italian and Spanish.

One striking outcome was the range of achievement across the participating jurisdictions, summarised in Figure 1 for the first foreign language (second foreign language was also addressed). While our experience as European citizens may have provided informal impressions of different countries' success in foreign languages, this was the first time that an empirical study was able to quantify levels of achievement, and the extent of the disparities were perhaps unexpected. It also made clear that countries' understandings of CEFR levels were widely divergent.

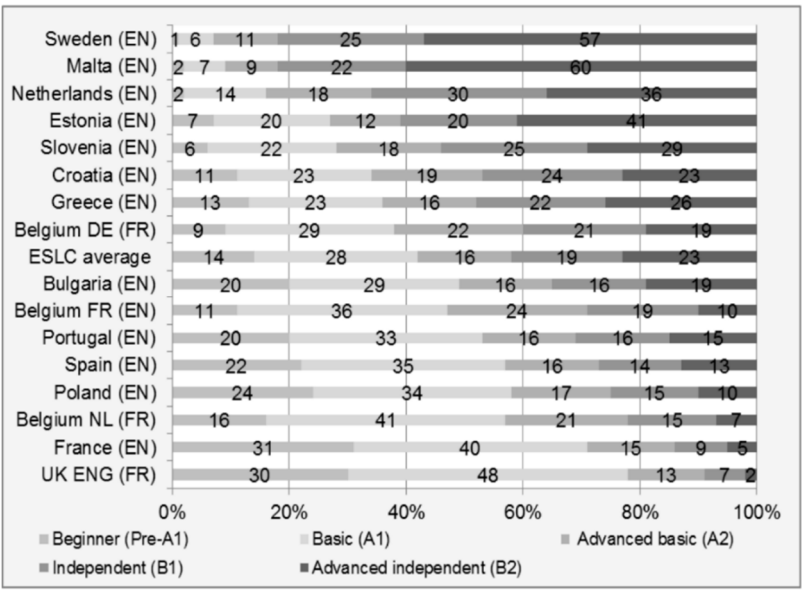

Figure 1: First foreign language. Percentage of pupils at each level by educational system (global average of the three skills) (European Commission 2012c: 9)

Figure 1 provides a summary league-table view of outcomes. Sweden comes top, with 57 percent of students achieving B2 in English, and England comes bottom, with 78% of students not doing better than A1 in French.

The ESLC is presented in a final report and a technical report, both available online (European Commission 2012a, 2012b). The purpose of this paper is to contribute to the theme of competence-oriented assessment, and will focus on the following issues:

- the status of the CEFR as a competence-based frame of reference,
- the test construction approach used in the European Survey to achieve comparability across languages,
- approaches to standard-setting,
- the potential of psychometric models to enhance the interpretation of complex data and to support the elaboration of useful comparative frameworks.

This paper also includes a brief account of the *Study on Comparability of Language Testing in Europe,* which the European Commission set in train in 2015 at the suggestion of the European Parliament. Rather than a repeat of the ESLC (countries had expressed some fatigue with educational surveys), they agreed to an alternative project, focused on evaluating the comparability of existing school language assessments, which, it was felt, might impact more usefully on educational processes in each country. The project, which was also delivered by Cambridge English, published its findings on 25th September 2015, the European Day of Languages. It is referred to here because it attempted to demonstrate scaling and standard-setting procedures which might engage countries much more directly, and which could contribute, in the belief of this author, to the aim of improving the comparability of educational standards across European jurisdictions.

2. The CEFR as a competence-based framework

There are several reasons for adopting the CEFR as the framework for reporting on levels of achievement in languages:

- As a familiar point of reference: The CEFR is widely referred to in Europe in relation to defining the goals of language education, professional training of teachers, curriculum development, and as a scale for reporting learning outcomes (even if there is considerable variation in how the CEFR levels are interpreted).
- As a relevant model of learning: Given its multiple authorship, the CEFR speaks with several voices on the nature of language learning, but at its centre is the action-oriented model which sees language skills developing through motivated interaction within society. This essentially social-constructivist, socio-cognitive model should have relevance, I will argue further below, to the language education goals of all jurisdictions.

– As a measurement construct: The second (and originally secondary) purpose of the CEFR is to offer a framework of levels, which sets out to enable a broad comparison of language learning programmes and purposes. Arguably it is as a framework of levels that the CEFR is best understood and most referred to (rightly or wrongly). Having provided the reporting scale for the first *European Survey on Language Competences* it is desirable that the CEFR should also be used to anchor further studies to the same scale. Going forward, we should see projects focused on the CEFR levels as potentially useful elements in a movement to bring national or regional language assessments progressively into better alignment.

The set of CEFR level descriptors A1 to C2, which have been widely adopted within Europe and beyond, represent a serious attempt to characterise progression in language learning through a behavioural scaling approach. Their author (North 2000) could defend the validity of the descriptors through their empirical calibration using an *Item Response Theory* (IRT) model. IRT is a psychometric approach which has revolutionised assessment in those contexts where it has been successfully adopted, and which points up the weakness of much educational assessment where it has not yet been adopted. I will introduce it in more detail in section 6 below. North could claim that the can-do scales which he contributed to the CEFR were more than subjective descriptive impressions: they reflected a shared understanding of progression in language competence as evinced by the analysed judgements of a large number of teachers.

More importantly, the CEFR levels have been adopted and further developed by several examination bodies, including Cambridge English Language Assessment and partners in the *Association of Language Testers in Europe* (ALTE). As will be pointed out below, this has contributed to the development of the CEFR, adding further dimensions for interpreting levels of achievement in language tests. Thus there are excellent models available for jurisdictions who wish to construct their own assessments and link them to the levels of the CEFR.

2.1 The CEFR's action-oriented model of use and learning

The text of the CEFR betrays its multiple authorship: looking in it for a view on the nature of language learning we will find a range of influences:

– the functional/notional approach of Wilkins (1976), also reflected in the *Waystage-Threshold-Vantage* series by van Ek and Trim (1990, 2001);
– the needs-analysis approach that follows from John Trim's work on a unit-credit system for adult learners;

- the behavioural scaling descriptive approach of Brian North's (2000) scales;
- a chapter on task-based learning;
- the notion of the action-oriented approach, which was contributed by Daniel Coste.

Of these it is the action-oriented model which most clearly reflects the social-constructivist position which I believe provides the most convincing model of how language learning proceeds.

There are two varieties of constructivism: *cognitive*, associated with Jean Piaget (Piaget 1976), and *social,* associated with Lev Vygotsky (Vygotsky 1986). These two constructivist positions – cognitive and social – are not at odds with each other, but can be seen rather as different emphases within an overarching concept of *situated cognition*, focusing either on the individual's cognition, or on the larger physical and social context of interactions and culturally constructed tools and meanings within which cognition develops.

Situated cognition captures the essence of learning: it happens in a social environment, it is purposeful and it is based in interaction. As the American philosopher, psychologist and educational reformer John Dewey (1859-1952) put it: "Education is a social process; education is growth; education is not preparation for life but is life itself" (1897). In Shepard's (2000: 12) presentation of social constructivism she comments that "John Dewey anticipated all of these ideas 100 years ago".

3. Constructs of Learning

The cursory treatment of cognition in the text of the CEFR was pointed out early by several scholars including Weir (2005a), who in collaboration with *Cambridge English* researchers went on to advance a *socio-cognitive* model of test validity Weir (2005b), which represents an essentially situated cognition viewpoint. Four construct volumes for Writing, Reading, Speaking and Listening (Shaw & Weir 2007; Khalifa & Weir 2009; Taylor 2011 and Geranpayeh & Taylor 2013) set out to supply the explicit construct models which the descriptor scales of the CEFR itself do not.

How constructs are implemented as test tasks is critical to the purpose of this paper, which is to follow the process from identifying the language skills to be tested, through test construction and administration, to interpreting the resulting response data. The figure below illustrates reading competence and shows a central cognitive core, accessed through strategies (on the left) and calling on a range of knowledge (on the right).

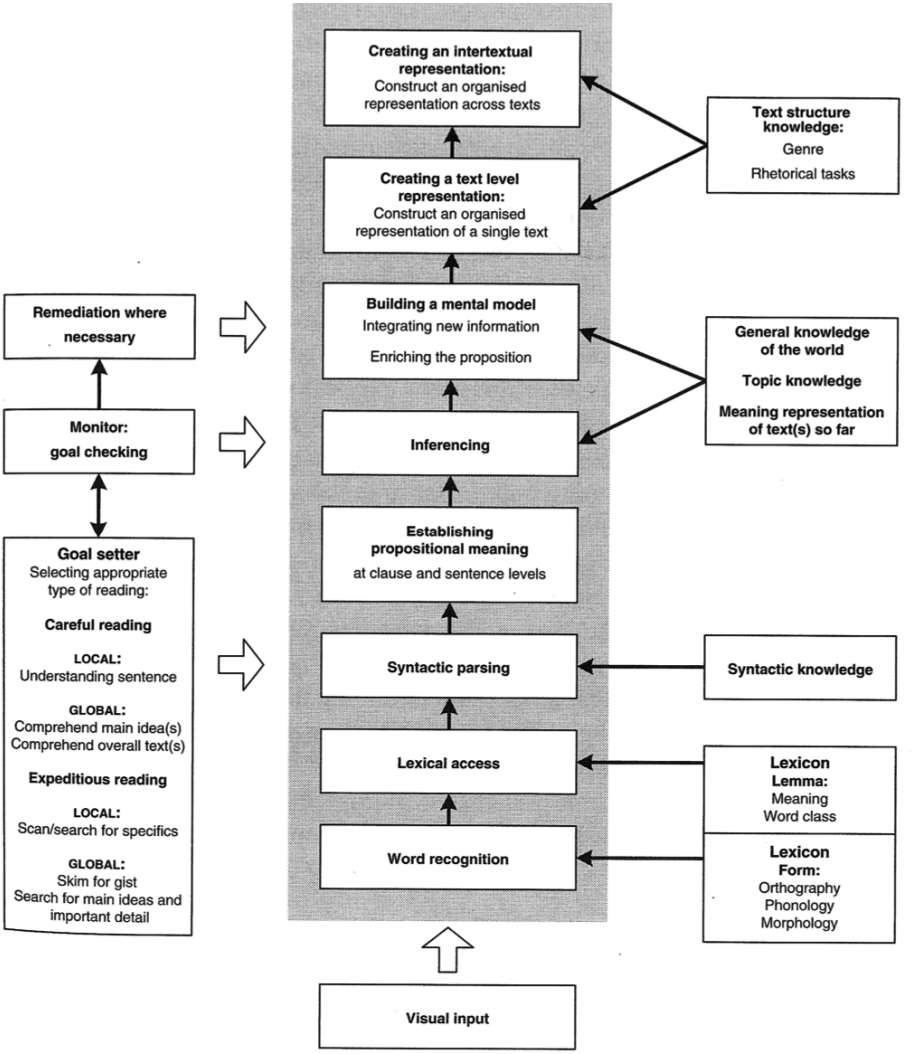

Figure 2: Illustration of a construct (Khalifa & Weir 2009: 43)

This model is based on relevant theory and supported by corpora of observed performance data. It is descriptive of how we believe cognition engages with reading, not prescriptive of how we believe it ought to. It is not a construct in the sense of a model arbitrarily constructed, but it is explicit about the posited cognitive processes, strategies and knowledge, and thus provides a good basis for setting item writers to work to construct test items which will be appropriate to a particular level of reading skill.

The construct represents our best understanding of learning and how it progresses for a particular skill. It will not fit every individual, precisely because language learning is situated in context; but it should do a reasonable job in providing the basis for a measurement scale. How that scale can be constructed is taken up in section 6 below.

3.1 What does "adopting the CEFR" mean?

As indicated above, the CEFR is a document of many parts. Of course, it is the descriptive scales provided by North (2000), implemented through a behavioural scaling approach, which most readers have focused on, and which define for them the meaning of the levels A1 to C2. The prominence of the descriptor scales is perhaps regrettable, particularly where it has led to unintended uses, such as the specification of curricula.

By developing the treatment of cognition an additional dimension of proficiency has been delineated, which complements the behavioural scaling approach, and is perhaps more appropriate to the aims of school assessment, focusing as it does on language in terms of cognitive development, rather than as a set of behaviours.

Above all the CEFR should be thought of not as a finished text but as an area of continuing work. When we speak of "adopting the CEFR" this need not imply uncritical acceptance of the current document: it has involved, and may continue to involve, further motivated development, including positive contributions by examination bodies. Another example of such development are the linguistic profiles developed for several languages, including the *English Profile* (2015), a significant corpus-based profile of linguistic features of English salient at each level of the CEFR (www.englishprofile.org), and *Profile Deutsch* (Glaboniat et al. 2005), which takes a slightly different approach to the description of CEFR proficiency levels in German. Through such work the CEFR levels have gained additional accretions of meaning and thus of value. Through such work the CEFR levels have gained additional accretions of meaning and thus of value.

4. The *European Survey on Language Competences*: test construction

The work at Cambridge on the constructs of language competence was a critical element of the approach to developing the language tests for the Survey, which reflected the CEFR's action-oriented, functional model of language use, while ensuring relevance for 14-17 year-olds in a school setting.

Five languages were specified for inclusion in the Survey, as the most widely taught in school: English, French, German, Italian and Spanish (the language partners were Cambridge English, the Centre international d'études pédagogiques (CIEP), the Goethe-Institut, the Università per Stranieri di Perugia and Instituto Cervantes with the Universidad de Salamanca). Of particular concern for the test developers, given the aim to develop comparable measures of achievement in five languages, was to work to a single test specification and to common item-writer guidelines, and ensure that all languages conformed to these. Each test construct was mapped to specific task types. The proportion of references to particular CEFR domains (personal, public, educational, professional) was specified for each CEFR level.

Most importantly the test developers, though they were located in five countries, worked as one team. They adopted a cross-language vetting system so that every test task was approved by the group as a whole. Many tasks were cloned across languages, particularly at the lower levels, where the relatively simple nature of tasks made this practical (care was taken to maintain the same cognitive challenge). As a result of this close collaboration across the five languages the final selection of tasks was highly comparable across languages.

At the pretesting phase tasks were progressively eliminated, so that the tasks used in the Survey were of very good quality from a psychometric point of view.

5. The *European Survey on Language Competences*: standard setting

Standard setting is the task which follows the construction of a measurement scale: it determines where on the scale the standards – CEFR levels in the case of the ESLC – should be set. Two approaches to standard setting were possible: it could be kept in-house, employing assessment expertise in the same closely-coordinated approach that had been used for test construction; or it could be opened to a wider range of participants. The perceived benefit of the latter approach was that the participating countries and jurisdictions could engage with the process, and in consequence might be more inclined to endorse the outcomes.

Additionally there was an argument that the standards were not for the consortium to dictate: rather, they could be seen to exist out there in Europe, with the consortium's task being to extract them from the combined judgments of as wide a range of participants as possible. This view won the day, and accordingly a large standard-setting event took place in Cambridge in September 2011. Panels varied in size from 21 for English to 8 for Italian (cf. European Commission 2012b)

The need for separate panels per language followed from the decision to invite a wider range of participants: it was not an option to restrict standard setting to a narrower group of multilingual experts. However, a cross-language alignment study on Writing, conducted online before the standard-setting conference, was able to verify that the standards set for Writing at least appeared reasonably comparable.

Standards were set for Listening, Reading and Writing. To describe the process for Listening and Reading: there were three rounds, which differed in their focus. Round 1 involved individual standard setting after having taken the test as a student. Round 2 involved individual standard setting after a discussion of the results of round 1. The purpose of this was to provide normative information – that is, to let individual members see how their judgments compared within the group of judges in terms of the standards set in the first round. The discussion was aimed at clarifying differences and to find out if these differences were due to either a misunderstanding of the CEFR or a misinterpretation of the demands of the tasks. It was expected that such discussion would lead to a decrease in inter-rater differences in the second round. Round 3 was intended as a validation procedure, described further below.

The conduct of the standard-setting activity illustrates an attempt to structure subjective judgments through an essentially psychometric view of competence. Judges were presented with sheets on which several tasks were ranked along a scale, representing the relative position of each item of each task. Judges were to indicate on the sheet the score for each test task which they believed represented the likely performance of a learner at the borderline between two levels. In the first two rounds there were three different sheets, each covering two overlapping levels between A1 and B2. The validation round presented all the test tasks on a single sheet (shown below).

Issues arose in the interpretation of the standard-setting data, as fully documented in the technical report (European Commission 2012b: 243 ff). It had proved possible for judges to make counter-intuitive judgments, with, for example, the order of B1 and A2 being reversed by some judges. The validation round also produced rather different results to the first two rounds (and indicated that the problem of reversed thresholds was linked to the use of separate sheets for judging). Thus the post-standard-setting process which had been planned to reconcile and finalise judgments turned out to be somewhat harder than expected.

In the event, the fastidious approach to test construction provided an argument for hypothesizing that the difficulty of the test tasks should be reasonably closely aligned across languages, and this assumption contributed to the reconciliation of the final standard-setting outcomes.

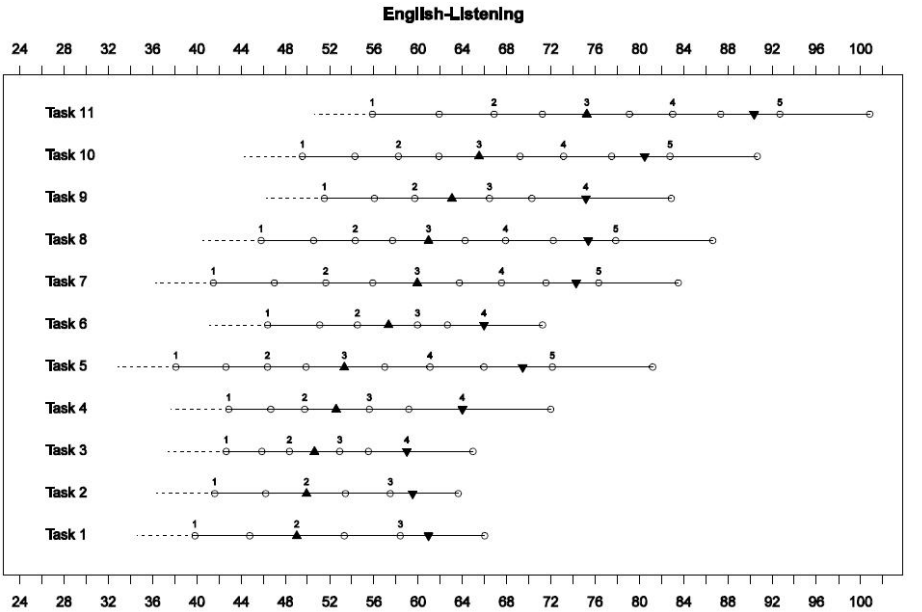

Figure 3: Example of the answer sheet for round 3 (European Commission 2012b: 282)

6. The measurement dimension

6.1 The metaphor of the trait

I have introduced in general terms the notion of scales of competence, upon which learners may be located according to their responses to test tasks. In this section I will look in a little more detail at IRT, the psychometric approach to constructing such scales.

Construct definitions provide a model of progression across levels. We treat competence in a particular skill as a linear *trait*. In language testing this simplifying assumption is generally moderated by the fact that exams test the 'four skills' separately, which allows individual profiles of skill to be captured. A student's exam outcome is then an average across the skills, which seems a defensible solution. The CEFR of course recommends reporting a profile rather than a

summary result, but both are useful; and the summary result seems appropriate to the situation where students study towards a CEFR level and may be judged as having achieved it or not, in some global sense.

Trait construction is supported by particular psychometric models, primarily based on IRT. The elegance of an IRT-based approach lies in the way that the resulting measurement scale co-locates the three essential elements of a testing situation: the *difficulty* of each task, the *ability* of each candidate and the *level* or cutoff-point for each grade awarded. Difficulty and ability are in fact mutually defining qualities. This potentially enables meaningful interpretation of performance, where each level can be characterised in terms of the group of tasks and the typical candidate behaviours elicited by those tasks.

A test which has been explicitly developed from construct definitions, as illustrated above, offers a straightforward measure of validity. The progression from lower to higher levels asserted by the construct model should be verified by the empirical item difficulties which emerge from the IRT analysis.

6.2 Item banking

The use of IRT is best illustrated on the case of item banking – the operational approach for constructing tests and interpreting test outcomes using IRT. The great value of item banking is that it creates an interpretive framework that can encompass exams at different levels, over different exam administrations and test versions, making it possible to generate tests with very similar measurement characteristics and to grade them to constant standards. Figure 4 gives a schematic view of item banking as a methodology for test construction.

The figure shows on the left an item bank containing tasks ready for use in a test. The difficulty of the items in each task is known, that is, they have been *calibrated*, using data from pretesting, and they are put on a single scale by using anchor tests, administered to candidates together with the pretests themselves.

Tests are then assembled using tasks of appropriate difficulty for the target levels. Candidates' scores on tests locate them on the measurement scale according to their ability. Figure 4 shows tests at three levels, and three candidates. Although they might all have the same score (say, 70%), we know that 70% on the easiest test indicates a lower ability than 70% on the hardest test: knowing the item difficulties enables us to locate the candidates precisely on the measurement scale.

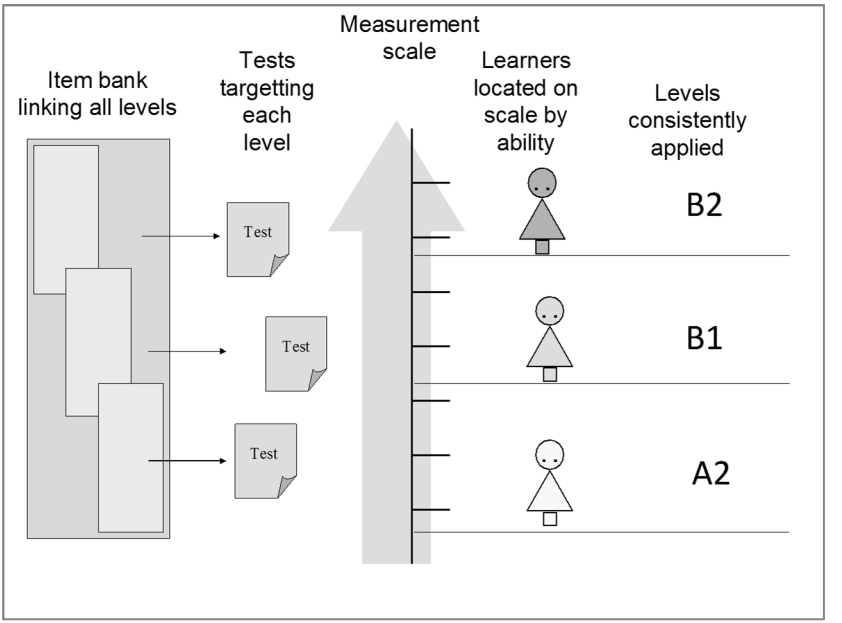

Figure 4: Item banking approach to scale construction and use (after Jones & Saville 2007: 501)

Finally, the standards are applied as fixed points which directly determine each candidate's grade. Even if test versions differ slightly in difficulty, the standard can be held constant. If we modify the standard it will impact all future tests in the same way. In such a fully-functional item banking system *ad hoc* standard setting is neither necessary nor possible.

Perhaps the most important benefit of an item banking approach is not simply that it facilitates the construction of a stable measurement scale, but that in consequence it facilitates the construction of *meanings* which explain what it is that the scale measures.

– Firstly, the items in the bank provide a concrete, detailed description of progression in terms of test content.

– Secondly, the fact that standards can be precisely maintained from session to session and from level to level facilitates doing the research to develop stable interpretations of learners' performance in the world beyond the test – for example in can-do statements such as those used in the descriptive scales of the CEFR.

– Thirdly, standards may described in linguistic terms. The *English Profile* (2015) is a large-scale study which has produced a linguistic description of CEFR levels, identifying salient features of data (Hawkins & Filipovic, 2012). All such developments exploit and contribute to the meanings embodied in the measurement scale.

In this way item banking enables *criterion-referenced* measurement, focusing on competences that link to real-world contexts of use.

6.3 Performance assessment

The objective testing of Reading and Listening using IRT shows it to be a technical and somewhat specialised approach to standardisation. However, the approach taken by large-scale assessment towards the performance skills of Speaking and Writing (at least in the practice of Cambridge) is more recognizable as a standardized version of activities that also take place in the classroom. Standardisation involves both judgments of performance and the nature of the performances themselves. Judgments are standardized by basing them on criterion-referenced exemplars, and by rating schemes which reflect as explicitly as possible the construct of Speaking or Writing at the targeted level. Training and monitoring of raters is an essential aspect of ensuring validity and reliability, and IRT can also be applied to standardizing raters' performance, by transparently compensating for the differences in severity which always exist.

7. The *Study on Comparability of Language Testing in Europe*

Relatively small in comparison with the effort placed by countries and contractors alike in the *European Survey on Language Competences*, this second study addressed a challenging but potentially game-changing conception: that existing national exam data might be used as a basis for making comparisons across jurisdictions. A study by Eurydice had provided a comprehensive list of language assessments and tests used by European countries (European Commission/EACEA/ Eurydice 2015). Countries were invited to provide material from these tests (focusing on a narrower range of languages, but including two ISCED (= *International Standard Classification of Level*)) levels for evaluation in the Survey. If it were possible to establish a basis for comparison, it might well focus educational research in Europe less on international surveys and league tables, and more on the effectiveness of current teaching and assessment practices – which are, after all, the prerequisites for success in the wider world.

The quantitative study used an approach known as comparative judgment, which is illustrated below. Comparative judgement provides an alternative source of data for an Item Response model: not using learner test response data, but rather a team of experts who perform the specific task of ranking samples of performance from better to worse. That is, they apply relative judgment, which is something humans are very good at, rather than absolute judgments of level, which humans find much more difficult (and which are often rooted in a specific local context). In this study Comparative Judgement was used to align samples of countries' test tasks to a common measurement scale, giving a picture of relative differences in difficulty. The scale was anchored to the CEFR by the use of anchor tasks taken from those used in the *European Survey on Language Competences*. The outcomes were interesting and suggested ways in which jurisdictions might with relatively simple means be enabled to develop a common evaluative framework based on psychometric techniques. Incidentally the Comparative Judgement approach would be capable of resolving the problem of pretesting for educational assessments, which in many jurisdictions is considered impossible for reasons of security. Of course, the devil is in the detail: countries' constructs of language competence may vary, for good reasons, and there are many differences in the way assessments are designed. Inevitably a review of reliability and validity shows up some weaknesses. Implementing the comparative approach would need refinement. Disappointingly, a lack of data prevented us proceeding to the step of linking countries' profiles of students' test results to the same CEFR-linked scale, via the difficulty of the tasks. Thus we were not able to demonstrate comparison of performance standards, although we could illustrate how it could be done.

The study was tasked to make several proposals: for how *post-hoc* adjustments could be made to increase the comparability of existing national results, for development work to increase the comparability of existing language tests, and guidelines for countries not currently using national assessments on how to develop new language examinations. It will be interesting to see countries' response to the study's final report, but it seems rather likely that the degree of coordination which any such developments would require will prove a decisive obstacle to further work in this area.

The Comparative Judgement approach used in the Survey was based on binary judgment: judges saw a series of two tasks, and rated one as the more difficult. Here I will illustrate a different approach to ranking (both achieve similar results). The multilingual benchmarking conference organised by CIEP at Sèvres in June 2008 focused on the performance skill of Speaking. It is interesting to report on here because two kinds of data were collected. At the conference itself judges rated video performances against the CEFR, with ratings elicited in a cascade design using English and French as anchor languages: working in one group (on

English and French), then in two and then three parallel subgroups, each dealing with three languages (i.e. English, French, and one other).

Prior to the conference ranking data were collected from the same judges, using a web-based platform which allowed them to view video samples and record their ranking by dragging samples to re-order them in a list. The allocation of samples for the ranking exercise was such as to ensure that each judge rated in two languages, and that there was linkage in the data across all samples and languages.

Figure 5 compares the abilities estimated from rankings and ratings for the set of samples submitted to both procedures. The correlation is high. Clearly there are some significant differences in the outcomes, but given that the ranking exercise took place before the conference, without guidance, discussion or familiarisation with the procedure, this is not surprising.

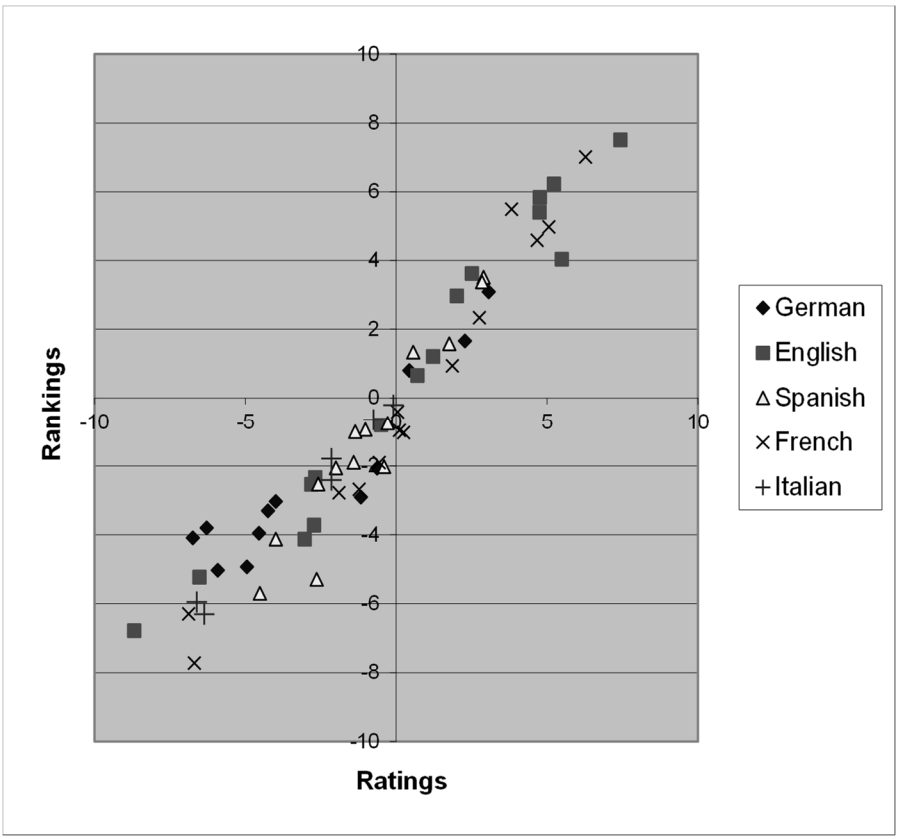

Figure 5: Ranking and rating compared (Breton 2008)

CEFR levels were assigned according to the judgments made at the rating conference, on the x-axis in the graphic – but could equally be transferred to the y-axis, so that new languages could be added easily in subsequent standard-setting exercises, by simply ranking the new examples against the existing ones.

Both the comparability studies briefly presented here illustrate relatively simple ways in which psychometric procedures can be brought to bear on organising and standardising human judgment in order to play to its strengths – that is, by making relative rather than absolute decisions.

8. Is my B1 your B1? Defining an approach to language education

In my presentation so far I have endorsed the value of the CEFR as a practical point of reference for language education. In this final section I would like to address the issues of interpretation and comparability which must be addressed in attempts to use the CEFR as a practical central point of reference.

As stated above, the CEFR is not a finished product: it is an area of ongoing work. This follows firstly from the CEFR's status as a frame of *reference:* every attempt to apply the CEFR to a particular context requires reflection and work to construct the link. Further, areas of research such as the language profiles or the Cambridge English work on cognition and constructs referred to above add more substance to the framework and provide more material for developing a meaningful link to a specific context.

It was Charles Alderson who in the early days of the CEFR asked the question: Is my B1 your B1? There are of course many ways in which one notion of B1 may differ from another:

- in the construct tested – the model of language competence
- in the learners to be described: young learners or adults and their respective cognitive styles, compulsory education contexts or language schools, etc.
- in the standard – which is harder, and which is "correct"?
- in the quality of measurement – we don't know whether my B1 is the same as yours if the tests are not reliable

Perhaps more important than any of these in the educational context is the view from the classroom: B1 may be an official goal of learning, but what are the skills which are taken to indicate its achievement? And what is the curricular content which produces the skills?

From the social-constructivist viewpoint taken in this paper it is important to distinguish curricular inputs from the criterion-referenced, real-world outcomes

of language study. The relation between the two is captured by the ontological concept of *emergence*: communicative competence is a higher-level *outcome* of learning which is qualitatively different to the curricular *inputs* to learning. The one cannot be derived directly from the other. Both must be addressed in the classroom. The concept of emergence asserts that the whole is more than the sum of the parts: a PPP (presentation – practice – performance) conception of teaching does not guarantee or explain the development of communicative competence.

The presentation of the CEFR above stressed the central importance of the action-oriented model of competence, asserting that it represents a relevant and widely applicable model of language learning. In this model communication is not just the goal – it is the prime mover. In the United Kingdom one may find skepticism regarding the concept of communication in language education, particularly in the tertiary sector, which seems to believe that the secondary sector has betrayed language education by "dumbing it down", reducing it to the teaching of phrasebook language. Perhaps there was a period when the novel concept of communicative language learning was indeed interpreted in this way. But if so that is to misunderstand what we intend by communication.

Communication is at the heart of the human condition. Shakespeare continues to communicate with us over four centuries, and the generation of learners growing up in the age of social media are finding new but still language-mediated ways of communicating and sharing their life experience. The natural desire to communicate is a powerful force for learning if it can only be harnessed.

The questionnaire findings reported in the *European Survey on Language Competences* support this contention. In contrast to the Commission's generally negative appraisal of the findings of the ESLC, Cambridge English provided the following positive interpretation, based on what the questionnaire responses clearly indicated:

> A language is learned better where motivation is high, where learners perceive it to be useful, and where it is indeed used outside school, for example in communicating over the internet, for watching TV, or travelling on holiday. Also, the more teachers and students use the language in class, the better it is learned. These conclusions only confirm what we already knew, but the European Survey provided empirical evidence in support of them. What the paragraph above describes is language being used for motivated, purposeful communication. However, the Survey shows that this ideal learning situation is approximated only in some countries, and effectively, only for English (Jones 2013).

From an assessment viewpoint, the construct of communicative language competence appears to be weakly developed in the teaching and testing regimes of many countries. An outcome of the *European Survey on Language Competences* was to demonstrate that countries' understandings of CEFR levels may vary widely: countries successful in language learning understand the levels as higher;

less successful countries understand them as lower – that is, levels are normed on those performances which are most familiar. But if it is accepted that a shared understanding of learning outcomes is a goal worth aiming at then there are practical ways of addressing it. I hope that the psychometric model which I have presented here, which takes us from the definition of constructs, through the stages of test design, test construction, standard setting and scale construction, will not be taken as a straitjacket on European language education, but rather as a practical approach to addressing important issues of how we define learning goals, shape inputs to learning, and compare learning outcomes.

Eingang des revidierten Manuskripts 10.12.2015

References

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)*. CIEP: Sèvres.

CEFR = Council of Europe (2001), *A Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: CUP [Online: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf, 28.01.2016].

Dewey, John (1897), "My Pedagogic Creed". *School Journal* 54 (January 1897), 77-80.

English Profile (2015), Cambridge: CUP (online: www.englishprofile.org)

European Commission (2005), *Commission Communication of 1 August 2005 – The European Indicator of Language Competence* [COM(2005) 356 final – Not published in the Official Journal], retrieved 18 January 2012 [Online: http://europa.eu/legislation_summaries/education_training_youth/lifelong_learning/c11083_en.htm, 28.01.2016].

European Commission (2007), *Terms of Reference: Tender no. 21: "European Survey on Language Competences"*. Contracting Authority: European Commission.

European Commission (2012a), *First European Survey on Language Competences: Final Report*. Luxembourg: Publications Office of the European Union [Online: http://ec.europa.eu/languages/policy/strategic-framework/documents/language-survey-final-report_en.pdf, 28.01.2016].

European Commission (2012b), *First European Survey on Language Competences: Technical Report*. Luxembourg: Publications Office of the European Union [Online: http://ec.europa.eu/languages/policy/strategic-framework/documents/language-survey-technical-report_en.pdf, 28.01.2016].

European Commission (2012c), *First European Survey on Language Competences: Executive Summary*. Luxembourg: Publications Office of the European Union [Online: http://ec.europa.eu/languages/library/studies/executive-summary-eslc_en.pdf, 28.01.2016].

European Commission/EACEA/Eurydice (2015), *National Tests in Languages in Europe 2014/15*. Luxembourg: Publications Office of the European Union.

Geranpayeh, Ardeshir, & Taylor, Lynda (eds.) (2013), *Examining Listening: Research and practice in assessing second language listening*. Studies in Language Testing Vol. 35. Cambridge: CUP.

Glaboniat, Manuela; Müller, Martin; Rusch, Paul; Schmitz, Helen & Wertenschlag, Lukas (2005), *Profile Deutsch* (5th edition). Berlin: Langenscheidt.

Hawkins, John A. & Filipović, Luna (2012), *Critical Features in English. Specifying the Reference Levels of the Common European Framework*. English Profile Studies Vol. 1. Cambridge: UCLES/Cambridge University Press.

Jones, Neil (2013), The European Survey on Language Competences and its significance for Cambridge English Language Assessment. *Research Notes* 52, 2-7.

Jones, Neil & Saville, Nick (2007), Scales and Frameworks. In: Spolsky, Bernard & Hult, Francis M. (eds.), *The Handbook of Educational Linguistics*. Oxford: Wiley-Blackwell, 495-509.

Khalifa, Hanan & Weir, Cyril J. (2009), *Examining reading: Research and practice in assessing second language reading*. Studies in Language Testing Vol. 29. Cambridge: Cambridge ESOL & Cambridge University Press.

North, Brian (2000), *The development of a common framework scale of language proficiency*. New York: Peter Lang.

Piaget, Jean (1976), *The grasp of consciousness*. Cambridge, MA: Harvard University Press.

Shaw, Stuart & Weir, Cyril J. (2007), *Examining Second Language Writing: Research and Practice*. Studies in Language Testing Vol. 26. Cambridge: Cambridge ESOL and Cambridge University Press.

Shepard, Lorrie A. (2000), The role of assessment in a learning culture. *Educational Researcher* 29: 7, 4-14.

Taylor, Lynda (ed.) (2011), *Examining Speaking: Research and practice in assessing second language speaking*. Studies in Language Testing Vol. 30. Cambridge: Cambridge ESOL and Cambridge University Press.

van Ek, Jan A. & Trim, John (1990, 2001), *Threshold 1990*. Strasbourg/Cambridge: Cambridge University Press, Council of Europe.

van Ek Jan A. & Trim, John (2001), *Vantage*. Strasbourg/Cambridge: Cambridge University Press, Council of Europe.

van Ek Jan A. & Trim, John (1990), *Waystage*. Strasbourg/Cambridge: Cambridge University Press, Council of Europe.

Vygotsky, Lev (1986), *Thought and language*. Cambridge, Mass.: MIT press.

Weir, Cyril J. (2005a), Limitations of the Council of Europe's Framework of reference (CEFR) in developing comparable examinations and tests. *Language Testing* 22: 3, 281-300.

Weir, Cyril J. (2005b), *Language Testing and Validation: An Evidence-Based Approach*. Oxford: Palgrave.

Wilkins, David A. (1976), *Notional syllabuses*. Oxford: Oxford University Press.