Research Papers in Language Teaching and Learning
Vol. 3, No. 1, February 2012, 8-21
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

The Landscape of Language Testing and Assessment in Europe: Developments and Challenges

[Γλωσσική Δοκιμασιολογία και Αξιολόγηση στην Ευρώπη: Εξελίξεις και Προκλήσεις]

Sauli Takala

The article opens with a short sketch of developments in language testing and assessment, and presents Spolsky´s tripartite categorization of major approaches in language testing/assessment. This is followed by an account of current developments in language testing and assessment in Europe. One prominent development in Europe has been a strong increase in cooperation in language testing/assessment. This was shown by the emergence of associations devoted to language testing/assessment, ALTE in 1990 and EALTA in 2004, both producing codes/guidelines for good practice. A major outcome of cooperation in the field of language education, the Common European Framework of Reference (CEFR; Council of Europe, 2001), brought about further challenges for language testing/assessment, immediately after its publication in 2001. Its potential was utilized by DIALANG, a pioneering internet-based system for diagnostic assessment and still apparently unrivalled. Most attention in the article is devoted to an analysis of future challenges. Justification of testing/assessment is increasingly challenged, and work is being done on analyzing assessment use argumentation. The problem of criterion, while not a new challenge, calls for attention both in terms of its logical status and various systems of standards published. Standard setting – setting one, or increasingly more often, multiple cut scores to indicate levels of proficiency is another current challenge, especially in efforts to relate tests, examinations etc to the CEFR levels. Following interest in international comparisons generated by the PISA programme, there is a trend to carry out international comparisons in the area of language competence (e.g. EILC).

ω

Το άρθρο ξεκινά με μια σύντομη επισκόπηση των εξελίξεων στο χώρο της γλωσσικής δοκιμασιολογίας και αξιολόγησης, και παρουσιάζει την τριμερή κατηγοριοποίηση των κυρίων προσεγγίσεων της γλωσσικής δοκιμασιολογίας και αξιολόγησης όπως προτείνεται από τον Spolsky. Στη συνέχεια ακολουθεί περιγραφή των τρεχουσών εξελίξεων της γλωσσικής δοκιμασιολογίας και αξιολόγησης στην Ευρώπη. Ένα σημαντικό χαρακτηριστικό της γλωσσικής δοκιμασιολογίας και αξιολόγησης στην Ευρώπη πρόσφατα είναι η ενδυνάμωση της συνεργασίας. Αυτό φαίνεται από την εμφάνιση οργανισμών γλωσσικής δοκιμασιολογίας και αξιολόγησης, όπως ο ALTE το 1990 και EALTA το 2004, οι οποίοι εξέδωσαν κωδικούς / οδηγίες για την ορθή πρακτική στο χώρο αυτό. Ένα σημαντικό αποτέλεσμα της συνεργασίας στον τομέα της γλωσσικής εκπαίδευσης, το Κοινό Ευρωπαϊκό Πλαίσιο Αναφοράς (ΚΕΠΑ - Council of Europe, 2001), επέφερε

περαιτέρω προκλήσεις στο χώρο της γλωσσικής δοκιμασιολογίας και αξιολόγησης, αμέσως μετά τη δημοσίευσή του το 2001. Το δυναμικό του ΚΕΠΑ χρησιμοποιήθηκε από το DIALANG, ένα πρωτοποριακό και ασυναγώνιστο σύστημα διαγνωστικής αξιολόγησης βασισμένο στο διαδίκτυο. Μεγάλο μέρος του παρόντος άρθρου είναι επίσης αφιερωμένο στην ανάλυση των μελλοντικών προκλήσεων. Η αιτιολόγηση της εγκυρότητας των δοκιμασιών και της αξιολόγησης όλο και περισσότερο αμφισβητείται, και περαιτέρω συζητήσεις περιστρέφονται γύρω από την ανάλυση της επιχειρηματολογίας της χρήσης των αξιολογητικών αποτελεσμάτων. Το πρόβλημα των κριτηρίων, ενώ δεν είναι μια νέα πρόκληση, απαιτεί προσοχή τόσο από την άποψη της λογικής κατάστασης του και τα διάφορα συστήματα κριτηρίων που δημοσιεύθηκαν. Η θέσπιση κριτηρίων που να ορίζουν το διαχωρισμό δύο ή περισσοτέρων επιπέδων γλωσσικής ικανότητας είναι μια άλλη σημερινή πρόκληση, ειδικά στο πλαίσιο των προσπαθειών να συνδεθούν τα διάφορα γλωσσικά κριτήρια, εξετάσεις, κλπ με τα επίπεδα του ΚΕΠΑ. Ακολουθώντας στενά τις εξελίξεις στο χώρο φαίνεται πως υπάρχει μια τάση για τη διενέργεια διεθνών συγκρίσεων στον τομέα της γλωσσικής ικανότητας (π.χ., EILC), η οποία δημιουργήθηκε από τις συγκρίσεις που υπέδειξε το πρόγραμμα PISA.

Key words: CEFR, standard setting, cut score, ALTE, EALTA, DIALANG, EILC, international assessment, criterion, IRT

Historical sketch

Spolsky (1995) presents a review of the history of language testing in his seminal work "Measured Words". He refers to the long history of testing and cites ancient China as a case where written high-stakes testing was used in civil service recruitment. Competitive examination made its way in various forms to European countries, and a modern variant is the competitive examination that the European Union arranges for all those who wish to become EU officials (linguists, administrators, assistants)[1]. Spolsky´s main focus was on language testing in Great Britain and the United States, comparing developments in language testing and their contexts. The main thread in his exploration and analysis is the development of objective language testing.

At the 1975 AILA conference in Stuttgart, Spolsky (1978) presented a much-quoted tripartite classification of periods or approaches in language testing and assessment: pre-scientific (later called "traditional"), psychometric-structuralist (later "modern") and psycholinguistic-sociolinguistic (later "post-modern"). He believes that there is a good deal of truth in the tripartite division but feels a bit uneasy as it was based more on impression than documented evidence. He suggests that it is, in fact, more useful to see the development of language testing as "an unresolved (and fundamentally unresolvable) tension between competing sets of forces" (Spolsky, 1995, 354). These forces are both practical and ideological.

There is also an early account of the development of language testing in the US by John B. Carroll, who by all accounts was a towering figure in language testing/assessment (and in measurement in general). His review covers the period from the late 1920s through to 1954, the year in which it was produced, but as it was never published, it is unfortunately little known[2]. The review covers 49 pages of texts, provides a comprehensive list of available tests and contains a bibliography of about 80 references.

At regular intervals – often at the entry of a new decade - there has been stock-taking in the form of congresses/seminars and related publications (conference proceeding, books). Some examples are: the papers presented at an international symposium in Hong Kong in 1982 (Lee et al., 1985); de Jong & Stevenson (1990); Huhta, Sajavaara & Takala (1993), which includes the plenary by Charles Alderson on the state of language testing in the 1990s, and Spolsky (2000), which is a review of articles on language testing/assessment published in The Modern Language Journal.

Current situation in language testing in Europe

Cooperation in language testing/assessment

No man is an island, and Europe is no island in language testing and assessment. Language testing/assessment has become an increasingly international domain of activity since the setting up of the Language Testing Research Colloquium, LTRC, (1979) and the International Language Testing Association, ILTA, (1992). Interest in specifically European co-operation in language testing and assessment is partly due to the influence of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), which was published in 2001 but was available for consultation and comments a few years before that. ALTE (Association of Language Testers in Europe) was founded in 1990, currently consisting of 34 members (examination providers) and covering 27 languages. EALTA (European Association for Language Testing and Assessment[3]) was created in 2004. In early 2011, it had 1232 Individual members, 145 Associate Members, 17 Expert Members and 54 Institutional Members. EALTA has intentionally aimed at a broad range of membership and inclusiveness, low costs and collegial support and co-operation.

Despite the growing interest in testing and assessment in Europe, American scholars have dominated the development of measurement theory, and they have also produced the most important tools and references, such as the handbook entitled "Educational Measurement" (Lindquist, 1951), which, since then, has appeared in four editions and the "Standards for Educational and Psychological Testing" (AERA, APA, NCME, 1999), which, since 1954, has appeared in six editions. Ethical and programme evaluation standards have also appeared. In these publications, concepts such as reliability, validity, fairness/bias have been addressed and continuously elaborated. However, it is Europe that has made one of the most important contributions to testing: in the early 1960s the Danish psychologist/mathematician Georg Rasch developed the powerful measurement model which goes under the name of Rasch modeling (Rasch, 1960; see also Verhelst, 2004).

The language testing associations have also contributed to awareness-raising about good practice. ILTA pioneered the work on ethical standards in language testing and assessment and both ALTE and EALTA have developed related guidelines. EALTA´s Guidelines for Good Practice in Language Testing and Assessment have been published in 34 languages.

The challenge posed by the CEFR for language testing and assessment

The Common European Framework of Reference (CEFR), drawing on decades of development work within the Council of Europe (CoE) and based on a decade of focused activity, was ready to be published in 2001. The approach to language teaching and learning promoted by the CoE, and largely subsequently adopted also by the European Union, was designed to be responsive to the needs of the increasingly cooperative political structures in our multilingual and multicultural continent. While the aim was communicative and intercultural competence, language projects were also expected to contribute to the basic CoE values of human rights, democratic citizenship, and rule of law. This meant, among other things, strengthening pluralistic participatory democracy, promotion of intensified international cooperation, understanding and tolerance of cultural and linguistic diversity as a source of mutual enrichment, and democratization of education, with languages for all (Trim, 2007). This orientation has even strengthened in recent times.
The CEFR was developed during the medium-term (1990-1997) project entitled "Language Learning for European Citizenship". It can be described in a number of ways depending on which of the rich content facets one wishes to stress. Trim (2007:39), one of the key architects of the CEFR, indicates its broad scope in a succinct characterization:

a) a descriptive scheme, presenting and exemplifying the parameters and categories needed to describe, first, what a language user has to *do* in order to communicate in its situational context, then the role of the *texts*, which carry the message from producer to receiver, then the underlying *competences* which enable a language user to perform acts of communication and finally the *strategies* which enable the language user to bring those competences to bear in action;
b) a survey of approaches to language learning and teaching, providing options for users to consider in relation to their existing practice;
c) a set of *scales* for describing proficiency in language use, both globally and in relation to the categories of the descriptive scheme as series of *levels*;
d) a discussion of the issues raised for curricular design in different educational contexts, with particular reference to the development of *plurilingualism* in the learner, and for the *assessment* of language proficiency and achievement.

While the broad scope of the CEFR is increasingly being recognised and appreciated, it is obvious that the proficiency scales and their use in testing, assessment and examinations have received most attention. The scales have been seen by decision-makers as a concrete means for defining learning targets and assessing learning outcomes. One aim has been to use them in comparing the performance of the national language teaching provision with other nations. As this appeared not to be an easy task, there were immediately calls for the CoE to undertake a validation or certificating function.

However, CoE´s mandate does not include such functions. Instead, in cooperation with the Finnish national authorities it organized a seminar in Helsinki in the summer of 2002, which led to the setting up of an international working group with the task of developing a manual to help in relating/aligning examinations and tests to the CEFR levels. A pilot manual was issued in late 2003 and a revised one, based on feedback received, in 2009. The manual presents five steps: familiarization with the CEFR, specification of the content of the tests/exams in accordance with the CEFR descriptive scheme, training and benchmarking with samples for oral and written performance, standardization of level judgments/ratings and validation of set cut-scores. A Reference Supplement edited by the present author was also produced to provide more technical information about the theoretical foundations of standard setting. The expanded version of the Reference Supplement is only available on the Council of Europe website (http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp).

DIALANG – pioneering and still unrivalled?

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

DIALANG[4] was in many ways a novel approach to language testing and assessment. It developed a transnational assessment system with a large range of languages covered. It is diagnostically oriented, with one of its goals to promote diversified language learning in Europe. It combines self-assessment and external assessment. It uses the Internet as the delivery system and reports the results in accordance with the Council of Europe proficiency scales. This linking was decided upon as the use of the scales was seen to promote comparability across languages. I played an active role in developing the blueprint and I recall seeing in the system an opportunity to “democratise” testing/assessment by trying to put the user “in the

driver's seat" and by serving his/her individual interests (to be "at his/her beck and call"). In this, DIALANG displayed a similar sense of "service mission" as EALTA.

In practice, it was the CEFR Draft 2 self-assessment scales and communicative activity scales that were used/adapted. We also reviewed and utilised the objectives and definitions in the CoE publications entitled Waystage (A2), Threshold (B1) and Vantage (B2). While we found these useful for test specification, we also noted that there was considerable overlap, and thus progression was not always clear-cut.

DIALANG faced many daunting challenges: how to write test specifications, self-assessment statements, feedback statements and relate all this to the CEFR (see Alderson 2005). Of course, relating the scores to the CEFR was a huge challenge (Kaftandjieva, Verhelst & Takala 2000) and it became a "hot" topic as soon as the CEFR had been published in 2001. It needs to be pointed out that standard setting in DIALANG required a new approach: from the usual task of setting one cut-score (failing/passing the standard), a situation which was then typical in the US, as many as five cut-scores were needed. This was done using the "modified Angoff" method as a starting point[5].

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

There has been intensive work on standard setting in language education, especially in Europe. Reference will only be made to a few recent major sources that standard setters should be aware of and consult: Figueras & Noijons (2009), Kaftandjieva (2010) and Martyniuk (2010).

Future Challenges

Justification of testing/assessment challenged

One of the challenges facing testing and assessment in the future is related to one of the main meanings of "challenge". As assessment literacy (i.e. awareness and competence in assessment) grows – even if one might wish to see a more rapid growth than is in evidence at the moment – it can be expected that the values underlying testing/assessment as well as its practices will be increasingly challenged. It is also probable that more openness and transparency will be demanded. Even if we Europeans often tend to criticize the excessive emphasis on testing and examinations in the US, it seems clear to me that they are ahead of Europe in accountability in testing. Testing/examination procedures can be, and are regularly, challenged in court. Thus testing/examination bodies know that they have to be able to present good evidence and arguments for their procedures, practices and decisions. Major evaluation/assessment studies are regularly analysed critically and challenged[6]. A good EALTA colleague, Dr. Felianka Kaftandjieva (who passed away in 2009), was deeply disturbed by what she saw as too common European "sloppiness" and lack of transparency in testing/examination accountability and she wished to see a challenging approach similar to the American one taking root in Europe. I am basically sympathetic to this view. Bachman and Palmer (2010) provide a good discussion of the need to elaborate a reasonable case for any assessment, that is, assessment use argument.

As a founding member of EALTA and its second President, I am very pleased that its Guidelines for Good Practice in Language Testing and Assessment, available in 34 languages, address a broad range of target groups, using simple and comprehensible language. In retrospect, I now believe that it would be useful to

add the decision makers to the groups addressed. Their actions influence all other groups and there is a great need for much better assessment literacy among them.

Closely related to the issue of accountability is the concern with the *ethics* of language testing. The consequences of assessment were singled out by Messick (1989, 1994) as an important aspect of the uniform concept of construct validity. Shohamy (2001) has provided a critical discussion of the power that tests exert in controlling people and institutions. International language testing and assessment associations have developed codes of ethics/guidelines of good practice.

The problem of the criterion

McNamara (2004) provides a clear and concise model which illustrates the relationship between the test, the construct and the criterion. In the model, the construct is placed in the middle, the test to the left and the criterion to the right. The construct represents the theory of the domain assessed and provides a description of the essential features of performance. It influences the operationalization of the construct, test development, and leads to observable responses to test items and to other kinds of performances. This observable data leads to inferences – via the theoretical model – about the real-world performance, about the testee's actual standing in relation to the domain. These are inferences about something that is unobservable as tests are always a sample from the overall domain.

The nature of the *criterion* is a perennial issue in language assessment. Davies (2003) has addressed the elusive but ubiquitous concept of the *native speaker* as a criterion in applied linguistics, SLA research and also in language assessment (cf. Abrahamson & Hyltenstam 2009). Davies concludes that the native speaker is a myth but a useful one. This criterion has, however, been increasingly questioned and the interest in English as a lingua franca (ELF) and "World Englishes" (Kachru, Kachru & Nelson, 2006) has further problematized this criterion.

A number of other criteria (standards) have been produced in language testing. Important contributions include the Foreign Service Institute's scales developed about 50 years ago for testing proficiency and subsequent scales of proficiency. The Common European Framework for Languages (CEFR; Council of Europe, 2001) developed in Europe has acquired a strong position, not only in Europe but also in some other parts of the world. Such a standards-based approach to assessment defines *content* as well as *performance standards*, usually called Performance Level Descriptors (PLD, see e.g. Cizek & Bunch, 2007: 44-47).

Self-assessment

The role of *self-assessment* has been increasingly recognized. Early cognitive psychology (e.g. Flower & Hayes, 1977) reinforced the view that effective learning consists of a number of processes of which important ones are the skills of planning and appraisal. An effective learner can plan how to approach the tasks, which requires an ability to evaluate one's current level in relation to the task, and the ability to monitor the process and to assess and – if need be – revise the output. The ability to assess one's language proficiency is seen as a powerful factor in language learning. Oscarson's (1980) early work on self-assessment within the CoE modern language project was seminal in promoting the concept. Recent research in Europe is reported by Dragemark-Oscarson (2009) and Huhta (2010) in their respective doctoral dissertations.

The European Language Portolio (ELP), closely linked with the CEFR, is an example of a tool making use of self-assessment. There are several versions of it ranging from early learning to higher education. It seems, however, that the promise of the portfolio has not been very extensively materialized.

Less attention, undeservedly, has been devoted to peer assessment, a potentially very useful approach to assessment.

Standard setting

Standard-setting (setting cut scores) is a challenge when assessments, tests and examinations increasingly are required to report the outcomes in terms of proficiency levels. Cizek and Bunch (2007) is a useful general reference. A Manual for Relating Language Examinations to the CEFR is also freely available (Council of Europe, 2009).

Jaeger (1989, p. 492) observed that "much early work on standard setting was based on the often unstated assumption that determination of a test standard parallels estimation of a population parameter. This means that there would be a correct value for the cut score". This would simplify things, but unfortunately, it is not true, as Zieky (2001, p. 45) observes: "there is general agreement now that cutscores are constructed, not found". Standard setting cannot be reduced to a problem of statistical estimation, to the proper use of the best psychometric methods. Zieky (2001, p. 46) notes that "clearly, the cutscore is what participants choose to make it. It is also now clear that what participants choose to make the cutscore depends on subjective values ... Setting a sensible cutscore requires a determination of which type of error is more harmful [masters fail to pass; non-masters pass]." This does not mean, as early critics of standard setting argued, that standard setting is arbitrary and a waste of time. It is subjective but it does not have to be arbitrary in the ordinary negative sense of the word. Camilli, Cizek & Lugg (2001, pp. 449-450) argue that standards can be considered acceptable if they follow a "psychometric due process", an analogy to legal practice.

Zieky (2001, pp. 29-30) lists a number of issues that, in his opinion, had not yet (in 2001) been clearly resolved despite some twenty years of research and development work[7].

- Which method of setting cuts cores will be the most defensible in a given situation?
- Exactly how should the participants in the standard setting job be trained? Does training have to be face to face? What is the minimum acceptable training time?
- What normative information, e.g. prior score distributions, should be given to participants, if any?
- When should the participants receive normative information?
- Should the participants receive information about item difficulty? If so, should they receive information about every item?
- Exactly what cognitive processes do the participants use (see e.g., Papageorgiou, 2010)?
- To what extent are the participants capable of making the required judgments?
- Should participants be told the likely effects of their judgments on the resulting cut score?
- Should any judgments be excluded from the calculation of the cut score? If so, what criteria should be used to exclude judgments?
- How should the standard error of measurement affect the cut score?
- How should variability in the participants' ratings affect the cut score?
- Should compromise methods be used that combine normative and absolute judgments? Which method is the most appropriate?

A monograph by Felianka Kaftandjieva (2010), which discusses general issues in standard setting and explores six methods in setting standards in relation to the CEFR, is probably the most comprehensive treatment on the topic so far. In addition to providing evidence of the quality of the methods, it gives useful recommendations for standard setting.

More general guidelines are also available. The highly influential *Standards for Educational and Psychological Testing* (AREA, APA & NCME, 1999) listed six standards applicable to setting cut scores in its 1985 edition. These had mainly to do with being aware of and reporting error rates (misclassifications) and

providing rationales and explanations. There were no clear standards about HOW to set cut scores. The 1999 *Standards* contain 10 standards, with obvious overlap with the previous standards, but also standards about the actual processes of setting cut scores.

The 1999 standards show a new emphasis on the *actual processes* of setting cut scores[8]:

- Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score (2.14).
- When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument (2.15).
- When raw scores are intended to be directly interpretable, their meanings, intended interpretations, and limitations should be described and justified in the same manner as is done for derived score scales (4.4).
- When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented (4.19).
- When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria (4.20).
- When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way (4.21).
- When statistical descriptions and analyses that provide evidence of the reliability of scores and the validity of their recommended interpretations are available, the information should be included in the test's documentation. When relevant for test interpretation, test documents ordinarily should include item level information, cut scores and configural rules, information about raw scores and derived scores, normative data, the standard errors of measurement, and a description of the procedures used to equate multiple forms (6.5).
- Publishers and scoring services that offer computer-generated interpretations of test scores should provide a summary of the evidence supporting the interpretations given (6.12).
- Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate that the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experience (13.6).
- If tests are to be used to make job classification decisions (e.g., the pattern of predictor scores will be used to make differential job assignments), evidence that scores are linked to different levels or likelihoods of success among jobs or job groups is needed (14.7).

International comparisons

There is a long tradition of empirical comparative research in educational outcomes. The pioneer was the IEA (International Association for the Assessment of Educational Achievement), which was started as a cooperative venture more than fifty years ago by a group of internationally minded educational researchers (e.g., Torsten Husén from Sweden and Benjamin S. Bloom from the US). In addition to international studies of mathematics and sciences, the IEA has carried out studies of English and French as a foreign language, of reading and literature (published in the early 1970s), and of writing in the late 1980s (the present author was the international coordinator). Studies of reading have continued and they have focused on 10–11-year olds (PIRLS, Progress in International Reading Literacy Study) in 2001, 2006 and 2011. OECD

(Organisation for Economic Co-operation and Development) initiated its PISA programme (Progress for International Student Achievement) in the early 2000s and has carried out three cycles of much publicized assessments (2001, 2006 and 2010).

Foreign language proficiency being one of the main priorities in the multilingual and multicultural Europe, EU recently initiated a project to assess the levels of proficiency attained. This initiative went under the name of "Barcelona Indicator" during the several years of preparations. A project was launched to explore the implications of this initiative called EBAFLS (Building a European Bank of Anchor Items for Foreign Language Skills[9]). It explored several questions, and one practical outcome was the infeasibility of using the L1 as the language of comprehension items in international testing. The project also found that many test items displayed strong DIF (differential item functioning, see Takala & Kaftandjieva, 2000) such that item difficulties tended to vary considerably among the participating countries. DIF is a major challenge and it will be very interesting to see the results of the EU project (EILC[10]) when they will become available in 2012.

In international comprehension tests (such as PISA), it is important to control the effect of the different L1 translations and to ascertain that the level of text and item difficulty is comparable across the contexts. In PISA reading comprehension (L1) the translated versions are based on a source text – usually either in English or French – which is translated into several L1 languages. There is a set of procedures to guarantee that the texts and the items are not only equivalent translations but also of equal difficulty. However, this is a great challenge, and the results have occasionally been challenged on this account and on account of suspected DIF. Hambleton and de Jong (2003) have addressed the issues and report on progress made. In two doctoral dissertations, colleagues at the University of Jyväskylä have addressed the issue of text authenticity in international reading literacy assessment (Sulkunen 2007) and the problem of translated text equivalence (Arffman 2007).

With the availability of DIF and other analysis options, the EU study is the first one facing the necessity of dealing with problems of international comparisons in L2. It is to be hoped that researchers will take an interest in carrying out more detailed analyses after the main results have been published. Past experience has indicated that there will be such a time pressure to publish the first results that more in-depth analyses cannot be reported at that stage.

Some psychometric challenges

This is a vast area and only a few points will be made.

There is a common wish to aim at *simple solutions* and usually it makes sense to start with simplified models. However, *language ability is a complex phenomenon*. In a monumental analysis of human cognitive abilities, Carroll (1993) identified 16 different abilities in the domain of language. We are justified to seek simplicity (practicality) in language assessment but we should be aware of our simplifications. Reckase (2010) notes that the traditional true-score theory (Classical Test Theory - CTT) and the more recent unidimensional Item Response Theory (IRT) give good approximations for some test situations. However, in other cases, more complex models are needed to accurately represent the relationships in the test data.
McDonald (1999) was probably the first to present a general introduction to test theory by introducing *items* as the starting point. This indicates the difference between the modern test theory (known as Item Response Theory) and Classical Test Theory, which focuses on test *scores*. Reckase (2010) considers this a good approach and he suggests that *items* are complicated. He also points out that they deserve careful attention as the quality of assessment is crucially dependent on the quality of items. This is a view that the present author has also espoused for quite some time and has found teaching courses about item writing very enjoyable. Training in item writing needs to be provided for teachers as well as for item writers who are commissioned to write items for examinations.

IRT-based psychometric models are being developed continuously. As far as I can judge from the literature, there is not full agreement on the soundness of their conceptual basis. Not having the competence to assess the merits of the arguments, I am happy to note that a recent contribution by Thomas Eckes (2010) on the many-facet Rasch measurement (MFRM) has been used to study the rating of composition. It is available in the Reference Supplement on the Council of Europe website. Those working on rating issues will benefit greatly from consulting the article.

There is also an increasing trend to use structural equation modeling, SEM, (including confirmatory factor analysis, CFA), e.g. in the study of the structure of motivation. There are some examples of the use of this approach also in assessment. A good example is Åberg-Bengtsson and Erickson (2006), which, among other things, is an interesting and sophisticated analysis of the internal structure of a Swedish national Grade 9 tests.

Conclusion

A perennial challenge (see Masters and Foster 2003) is to guard against placing undue emphasis on a limited range of easily-measured skills at the expense of students' abilities to apply their learning, to reflect on and think about what they are learning, and to engage in higher-order activities such as critical analysis and problem solving. One concrete aspect of this in language testing and assessment is to reflect whether we should, for instance, focus only on testing comprehension of text or also give some attention to learning from text. What would such a test look like?

A similarly persistent challenge is to ascertain where all students are in their learning. If a broad range (say, three CEFR levels) should be distinguished in a test/exam, the assessment tasks need to provide a challenge to, and yield useful information about an equally broad range of proficiency. Valued learning outcomes may require the use of assessment methods which are not common in large-scale assessments/examinations. In principle, this approach is relatively easy to implement/carry out in classroom assessment: direct observations and judgments of students' work and performances over time (e.g. in ELP). Masters and Forster (2003) suggest that having a sufficiently broad coverage of valued outcomes may involve greater use of open-ended tasks that allow students to respond at a variety of levels, or tests that do not require all students to attempt exactly the same items (e.g., tailored tests in which students take items of different difficulty). All this sets high demands on task construction and also on reliability (for two cut scores a reliability of at least .941 is required; Kaftandjieva 2010).

In a recent book, which provides an interesting discussion of assessment use argumentation, Bachman and Palmer (2010) note that there are frequent *misconceptions and unrealistic expectations* about what language assessments can do and what they should be like. They list (a) a misguided belief in one "best" way to test language ability for any given situation; (b) a belief that language test development depends on highly technical procedures and should be left to experts; (c) a belief that a test is either "good" or "bad", depending on whether it satisfies *one particular* quality instead of a number of requisite qualities.

Socio-cultural theory (SCT) is probably a very good candidate to contest views of what is good and bad in current practices in language testing and assessment. In a recent book, Swain and her colleagues (Swain, Kinnear & Steinman 2011) devote a chapter to assessment. They list the following as SCT tenets related to second/foreign language assessment:

- assessment is social and cultural activity
- language performance is co-constructed
- language instruction and language assessment form a dialectical unity of language development

- fairness and equity in language assessment occur during ZPD (Zone of Proximal Development, a concept introduced by the Russian psychologist Vygotsky to indicate the developmental stage that the learner is approaching)

The authors rightly suggest that SCT-inspired assessment will challenge traditional assessment approaches – and create controversies – about validity, reliability, scoring and fairness. Much work is needed to deal adequately with the controversies that are likely to emerge. Kane (2006) will be an indispensible source in this work.
In the European assessment context it is necessary to refer to the extensive and noteworthy work carried out by the Assessment Reform Group (e.g., Harlen 2007, Stobart, 2008).

It also goes without saying that computer adaptive testing will develop and will offer new opportunities as well as challenges.

In conclusion, I wish to cite Davies (2003) who has cautioned about "heresies" of languages testing research, resulting from too enthusiastic embracing of new approaches and leading to loss of proper balance. Therefore, developing *assessment literacy* in a wide sense is a permanent challenge in language testing and assessment. Davies cautions about heresies but he also welcomes them as an antidote to moribund orthodoxy. This is reminiscent of a dictum by Alfred North Whitehead, Russel's teacher and co-author and an endless source of challenging quotes, to the effect that "wherever there is a creed, there is a heretic round the corner or in his grave" (Whitehead, 1993, p. 52).

Author's e-mail: sjtakala@hotmail.com

Notes

1. See e.g. http://europe.eu/epso/index_en.htm.
2. Another little known but very useful early review of the profession is by Stern, Wesche & Harley (1978).
3. www.ealta.eu.org.
4. This section on DIALANG is based on my presentation at the joint IATEFL-EALTA conference on "Putting the CEFR to Good Use", Barcelona, October 29-30, 2010.
5. Actually three different modifications of the modified two-choice Angoff method as well as three different modifications of the contrasting group-method were applied to the standard setting procedure. Multiple matrix sampling with incomplete equal-sized linked design was used to pilot the items. Item response theory was applied to item calibration. The One Parameter Logistic Model (OPLM) was chosen, because it combines the desirable statistical characteristics of the Rasch model with the attractive features of the two-parameter logistic model. Moreover, the OPLM computer program allows application of incomplete test design, which at that time was not possible with most of the other computer programs that applied the IRT approach to test development and analysis. The adaptive test construction design was based on the two-stage multilevel adaptive testing approach. The role of the routing test (pre-estimation) is played by the Vocabulary Size Placement Test and the self-assessment tools. The second-stage language test has three overlapping levels of difficulty. For standard setting, see Cizek & Bunch (2007).
6. See e.g. National Education Policy Center, http://nepc.colorado.edu.
7. It can be claimed that, in spite ten more years of R & R on standard setting, Zieky´s questions have not received a definitive answer and new issues actually emerge all the time.
8. These are cited in full (but leaving out the annotated comments) as standard setting in language testing and assessment is such a topical question in Europe but there seems to be too limited awareness of the *general standards* that apply also in language testing. Living up to these standards is a huge challenge in language testing and assessment.
9. http://cito.com/research_and_development/participation_international_research/ebafls.aspx.
10. http://europa.eu/legislation_summaries/education_training_youth/lifelong_learning/c11083_en.htm.

References

Åberg-Begtsson, L. & Erickson, G. (2006). ´Dimensions of national test performance: A two-level approach.´ *Educational Research and Evaluation*, 12(5), 469-488.

Abrahamson, N. and Hyltenstam, K. (2009). 'Age of acquisition and nativelikeness in a second language: listener perception versus linguistic scrutiny', *Language Learning,* 59/2: 249-306.

AERA, APA, NCME, (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alderson, J.C. (2005*). Diagnosing foreign language proficiency. the interface between learning and assessment.* London: Continuum.

Arffman, I. (2007). *The problem of equivalence in translating texts in international reading literacy studies. a text analytic study of three English and Finnish texts used in the PISA 2000 reading* test. Jyväskylä: University of Jyväskylä Institute for Educational Research.

Bachman, L. and Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Camilli, G., Cizek, G.J. & Lugg, C. A. (2001). ´Psychometric theory and the validation of performance standards: history and future perspectives.´ In G. J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives*. Mahwah. N.J.: Erlbaum, 445-476.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting. a guide to establishing and evaluating performance standards on tests.* Thousand Oaks: Sage.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2009). 'Relating language examinations to the Common European Framework of Reference for anguages: learning, teaching, assessment: a manual.' Accessed at http://www.coe.int/T/DG4/Portfolio/documents/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf on 23 Nov 2011.

Davies, A. (2003). *The native speaker: myth and reality*. Clevedon: Multilingual Matters.

de Jong, J. H. A. L. & Stevenson, D. K. (Eds.) (1990). *Individualizing the assessment of language abilities*. Clevedon: Multilingual Matters.

Douglas, D. & Chapelle, C. (Eds.) (1993). *A new decade of language testing research*. Alexandria: TESOL.

Dragemark-Oscarson, A. (2009). *Self-assessment of writing in learning English as a foreign language: a study at the upper secondary school*. Gothenburg: University of Gothenburg.

Eckes, T. (2010). 'Many-facet Rasch measurement. Section H of the reference supplement to the preliminary version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment', accessed at http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf on 11 Nov 2011.

Lindquist, E. F. (Ed) (1951). *Educational measurement*. Washington, D.C.: American Council on Education.

Figueras, N. Noijons, J. (Eds) (2009). *Linking to the CEFR levels: research perspectives.* Arnhem: Cito/EALTA.

Flower, L. & Hayes, J.B. (1977). ´Problem-solving strategies and the writing process.´ *College English* 39/4: 449-461.

Hambleton, R. K. & de Jong, J.H.A.L. (2003). ´Advances in translating and adapting educational and psychological tests.´ *Language Testing*, 20/2: 127-134.

Harlen, W. (2007). *Assessment of learning*. Los Angeles: Sage.

Hughes, A. & Porter, D. (Eds.) (1983). *Current developments in language testing*. London: Academic Press.

Huhta, A., Sajavaara, K. & Takala, S. (Eds) (1993). *Language testing: new openings.* University of Jyväskylä: Institute for Educational Research.

Huhta, A. (2010). 'Innovations in diagnostic assessment and feedback: An analysis of the usefulness of the DIALANG language assessment system.' Unpublished PhD thesis, University of Jyväskylä.

Jaeger, R.M. (1989). ´Certification of student competence.´ In R. I. Linn (Ed.), *Educational measurement. (3rd ed.)*. Washington, D.C.: American Council on Education, 485-514.

Kachru,B. Kachru, Y & Nelson, C. (Eds) (2006). *The handbook of World Englishes*. Malden, MA: Blackwell Publishing.

Kaftandjieva, F. (2010). 'Methods for setting cut scores in criterion-referenced achievement tests: a comparative analysis of six recent methods with an application to tests of reading in EFL.' Accessed at www.ealta-eu.org/resources on 23 Nov 2011.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In J. C. Alderson (Ed.), *Common European Framework of Reference for languages: learning, teaching, assessment: case studies.* Strasbourg: Council of Europe, 105-129.

Kaftandjieva, F., Verhelst, N. & Takala, S. (2000). 'Standard setting procedure'. Unpublished manuscript.

Kane, M. T. (2006). 'Validation´. In R. L. Brennan (Ed.), *Educational measurement (4th ed)*. Westport, CT: American Council on Education, 17-64.

Lee, Y. P., Fok, A. C. Y. Y., Lord, R. & Low G. (Eds.) (1985). *New directions in language testing.* Oxford: Pergamon Press.

Masters, G. & Forster, M. (2003). 'The assessments we need.' Accessed at http://cunningham.acer.edu.au/inted/theassessmentsweneed.pdf on 11 Nov 2011.

McDonald, R. M. (1999). *Test theory: a unified treatment.* Mahwah, N.J.: Erlbaum.

Martyniuk, W. (Ed.) (2010). *Aligning tests with the CEFR: reflections on using the Council of Europe's draft manual*. Cambridge: Cambridge University Press.

McNamara, T. (2004). ´Language testing.´ In A. Davies and C. Elder (Eds.), *The handbook of applied linguistics*. Oxford: Blackwell, 763-783.

Messick, S. (1989). 'Validity'. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. New York: American Council on Education & MacMillan, 13-103.

Messick, S. (1994). ´The interplay of evidence and consequences in the validation of performance assessments.´ *Educational Researcher*, 23: 13-23

Oscarson, M. (1980). *Approaches to self-assessment in foreign language learning*. Oxford: Pergamon Press.

Papageorgiou, S. (2010). 'Investigating the decision-making process of standard setting participants'. *Language Testing,* 27/2: 261-282.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Reckase, M. D. (2010). ´NCME 2009 presidential address: what I think I know.´ *Educational Measurement: Issues and Practice*, 29/3: 3-7.

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

Spolsky, B. (1978). ´Language testing as art and science.´ In G. Nickel (Ed.), *Proceedings of the fourth international congress of applied linguistics, vol. 3*. Stuttgart: Hochshulverlag, 9-28.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Spolsky, B. (2000). ´Language testing in the Modern Language Journal.´ *Modern Language Journal*, 84/4: 536-552.

Stern, H. H., Wesche, M. B. & Harley. B. (1978). The impact of the language sciences on second-language education. In P. Suppes (Ed.), *Impact of research on education: some case studies*. Washington, D.C.: National Academy of Education, 397-475.

Stobart, G. (2008). *Testing times. the uses and abuses of assessment*. London: Routledge.

Sulkunen, S. (2007*). Text authenticity in international reading literacy assessment: focusing on PISA 2000*. Jyväskylä: University of Jyväskylä.

Swain, M., Kinnear, P. & Steinman, L. (2011). *Sociocultural theory in second language education: an introduction through narratives.* Bristol: Multilingual Matters.

Takala, S., & Kaftandjieva, F. (2000). 'Test fairness: A DIF analysis of an L2 vocabulary test'. *Language Testing*, 17/3: 323-340.

Trim, J.L.M. (2007). 'Modern Languages in the Council of Europe 1954-1997: International co-operation in support of lifelong language learning for effective communication, mutual cultural enrichment and democratic citizenship in Europe', accessed at http://www.coe.int/t/dg4/linguistic/Source/TRIM_21janv2007_%20EN.doc on 11 Nov 2009.

Verhelst, N. (2004). 'Item response theory. Section G in the Reference Supplement to the Pilot version of the Manual for relating language examinations to the Common European Framework of Reference for languages: learning, teaching, assessment', accessed at http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionG.pdf on 23 Nov 2011.

Whitehead, A. N. (1933). *Adventures of ideas.* New York: New American.

Zieky, M. J. (2001). ´So much has changed: how the setting of cut scores has evolved since the 1980s'. In G. J. Cizek (Ed.) *Setting performance standards: concepts, methods, and perspectives*. Mahwah, N.J.: Erlbaum, 19-51.