EDITORIAL

Open Access

Foreword to the Special Issue "The Common European Framework of Reference for Languages (CEFR) for English Language Assessment in China" of *Language Testing in Asia*

J. Charles Alderson

Correspondence:
c.alderson@lancaster.ac.uk
Department of Linguistics and English Language, Lancaster University, Lancaster, UK

Background

The need within language testing and assessment for specifications of test content, methods and constructs has long been recognised and accepted. In education more generally, statements of aims, objectives and curricular frameworks are also widely provided, although the definitions and operationalisations of these may vary greatly.

In second and foreign language education, there is a long tradition of stating expected levels of achievement, with or without reference to curricular objectives. However, such stated levels were often vague and were not defined independently of higher or lower levels. Hence, it was very common to use terms like Beginner, False Beginner, High or Low Intermediate and Advanced or Mastery, the meanings of which were interpreted very differently in different contexts, cultures and educational systems, such that one person's or system's "False Beginner" might be another's "Intermediate".

More recently, efforts have been made to define achievement levels in terms of the requirements of educational contexts or of learners' needs and their ability to function independently in particular settings (see, for example, the Threshold level developed by language specialists under the aegis of the Council of Europe in the 1970s or the Vantage and Mastery levels in the 1980s). Draft versions of the Common European Framework of Reference were developed in the 1990s, which speculated on or stated what learners at particular levels could do with the particular language being assessed.

The DIALANG suite of diagnostic language tests in 14 European languages (https://dialangweb.lancaster.ac.uk/) was based on a draft (1996) version of the Common European Framework of Reference for Languages (CEFR) and used Can Do statements taken from the CEFR as its basis for test item development and the reporting of test results. Partly as a result of the success and take-up of DIALANG, the CEFR was rapidly adopted in many European educational systems, and CEFR-based exit levels from secondary or higher education were required by law in some countries (e.g. the Austrian Ministry of Education and the University of Innsbruck as reported in Spöttl et al. 2016).

This requirement to use the CEFR was not without controversy (see Fulcher, 2004, for example, or Weir, 2005), but their arguments were largely ignored by language educators. Language policies were widely developed with the CEFR at their core, both within and beyond Europe, resulting not only in CEFR-based language tests, but also CEFR-based language textbooks and language education curricula. Despite, or perhaps because of, the wide and rapid adoption of the CEFR, the somewhat inevitable limitations of the CEFR became apparent, and these are described in the paper by Jin, Wu, Alderson and Song in this collection of papers.

As a result of these perceived limitations, Japan developed its own localised version of the CEFR, the CEFR-J (see Jin et al.). More recently, the Chinese Government, aware of the existence, nature and impact and spread of the CEFR, decided to develop its own national framework and standards for English. The National Education Examinations Authority (NEEA) was charged with developing a theoretical model for a national framework (NEEA, 2014), which quickly became known as the China's Standards of English (CSE).

Questions that readers may ask themselves when reading this Special Issue include How does the CEFR influence the CSE, or perhaps, How does the CSE draw on and use/diverge from/improve the CEFR? What evidence (empirical and theoretical) for validity is presented or needed in the CSE? What data is available on the reliability and levels of descriptors? What information is available in the various papers on the actual content of scales, test types, test features, test development and design? What are the issues in localisation and prescription of the CSE across China? How are the tests to be developed based on the CSE, and how will they be implemented and administered in practice? What problems emerge and how are they addressed? What impact and washback will CSE-related tests have?

Developing the China Standards of English: challenges at macropolitical and micropolitical levels

The paper by Jin et al. is a key document that will repay careful attention and re-reading, as it makes the case for a national framework for English education in China. It spells out the need for such a framework, the potential benefits of the China's Standards of English, the challenges of such a project and potential pitfalls. It provides a critique of the Common European Framework of Reference (CEFR) and argues for a national rather than an international framework. It acknowledges that such an ambitious project will encounter resistance and criticisms from vested interests and foresees that macro-politics as well as micro-politics will play an important role in this.

The paper is structured around four key questions:

> Why do we need a national framework of reference for English in China?
> Why do we create a new framework instead of adopting or adapting an existing one?
> What are the challenges facing government organisations in the process of constructing and implementing a national framework?
> What are the challenges facing the individuals involved in or having stakes in the construction and development of a national framework?

It is hoped that a national framework will provide a shared understanding of the standards of English education and assessment, that it will also have a positive impact on the quality of teaching, learning and assessment of English and that it "will better prepare Chinese people to live and work in an increasingly globalised world".

One may legitimately wonder whether the CSE project is re-inventing the wheel and whether, after all the effort and resources going into it have been expended, the resulting CSE will be significantly different from the CEFR. Only time will tell, but this is a brave attempt to innovate in what will certainly involve difficult circumstances in such a huge and varied country. Differences between the Chinese and the European contexts are acknowledged, and some of the limitations of the CEFR are discussed and will hopefully be avoided. The authors of this lead paper remain positive but acknowledge that the project will need to continue to develop, be revised and followed up with a range of accompanying developmental and research projects.

Developing common listening ability scales for Chinese learners of English

The paper by He and Chen describes progress in the development of a set of scales for Listening in a second or foreign language. This will certainly have to be referred to in any similar future project not only in listening but also in other language skills, including Reading of course, but also Grammar and Mediation at the very least.

The basic constructs investigated are cognitive ability, listening strategies, linguistic knowledge and performance in typical listening activities. Cognitive ability is to be assessed by scales of narration, description, exposition, argumentation, instruction and interaction. Listening strategy is addressed by scales of planning, execution and evaluation and repair. Typical listening activities will involve scales of listening to conversations, listening to lectures, listening to announcements and instructions, listening to broadcasts and watching movies and TV series. Use of linguistic knowledge will involve scales of grammar and pragmatics. Interestingly, vocabulary is not addressed in this paper on listening but in a separate paper in the collection by Zhao, Wang, Coniam and Xie.

This article reports a descriptive approach to scale development, which presented a number of difficulties that have yet to be resolved. The four research questions that were formulated were:

> How do we define the construct of listening ability with respect to the English teaching and learning context of China?
>
> How do we describe listening ability in a comprehensive manner?
>
> How do we collect descriptors with reasonable representativeness?
>
> How do we scale the descriptors and how to validate the scales?

The data collection itself was extensive. The researchers examined 42 documents in both English and Chinese, including proficiency scales, teaching syllabuses, curriculum requirements, test specifications and rating scales, resulting in 1240 descriptors. They invited 159 teachers, 475 students and 119 professionals from different fields to write descriptors, which resulted in 1263 descriptors. Nationwide large-scale questionnaire surveys are planned to be distributed to 10,000 teachers and 100,000 students in over 1000 schools.

I look forward to reading the results of this ongoing research in due course.

Developing reading proficiency scales for EFL learners in China

The paper by Zeng and Fan on developing scales for Reading is in many ways parallel to the paper on Listening, since they both deal with language comprehension, but in other ways, it is different, especially in the research questions asked. The authors give a very brief account of the growing number of proficiency scales around the world, while pointing out that China does not have nationwide proficiency scales. They introduce the CEFR which is being drawn upon for the development of the CSE but argue that there is a need for localisation of the CEFR which would better reflect the current practice of learning and teaching of English in China as well as having positive effects on English learning and teaching.

Two questions are addressed in the paper, namely "What is the theoretical basis for developing reading scales?" and "What are the parameters for describing reading proficiency?"

Core members of the project designed the methodology and trained teachers to compile and develop reading descriptors. They were also responsible for collecting relevant literature and analysing interviews with experts and the descriptors developed by a working group consisting of 12 teacher leaders, 94 other teachers and a team of experts in language testing or second language education. Useful details are given of the documents analysed (which are largely similar to those that the original developers of the CEFR also consulted, with the addition of some Chinese documents) and the procedures developed for drafting descriptors and for refining the parameters for describing reading proficiency. Figure 1 gives a useful overview of this process.

Similar to the Listening descriptor development team, the Reading project members developed a framework for the CSE Reading scales consisting of Cognitive Ability, Comprehension Strategy and Knowledge, and some details are given of the components of these constructs. Similar to the development of the Listening descriptors, a large number (14,467) of reading descriptors was gathered, of which 1398 were summarised by team members from the literature and 13,069 were compiled by teachers. Again, useful details are given of the process of refining the parameters for classification of the descriptors and for revising and removing those draft descriptors that did not conform to these parameters. This process initially resulted in 4884 descriptors which were further reduced to a total of 574 descriptors to be entered in the database. Finally, the paper briefly discusses the improvement of the theoretical basis of the resulting scales, the removal of references to native-speaker standards and norms for reading proficiency (which also happened in the development of the CEFR scales) and the grouping of the descriptors according to the function intended by each descriptor.

This paper gives a fairly detailed outline of the theoretical basis for the procedures, constructs and parameters for developing reading descriptors, without, however, much practical exemplification of the results. Such exemplification will eventually have to be achieved in future research and development by showing how and why descriptors were accepted, rejected or revised during the process outlined in this paper. This research should involve both qualitative and quantitative empirical studies, interviews and questionnaire surveys and analyses and standard-setting studies as mentioned below in order to identify proficiency levels and cut scores. Such research will certainly need to extend well beyond the currently envisaged deadline of 2017.

Looking beyond scores: validating a CEFR-based university speaking assessment in Mainland China

Unlike other papers in this Special Issue, the article by Liu and Jia does not look at the development of descriptors, scales and test constructs but rather seeks to validate a localised university-based speaking test which is based on the CEFR. In contrast to the paper by He and Chen, this is a small study involving 54 learners, two interlocutors and two raters. Moreover, the researchers analyse transcriptions of videoed performances of only one part of the speaking assessment, with first year university students. Unusually, the researchers concentrated on counting the language functions displayed by the students and explored which features of speaking distinguish speaking proficiency at each level. In the end, however, it proved impossible to relate functions and features like length of turns, choice of words and syntax, hesitation markers and topic coherence to speakers' level of oral proficiency.

This is nevertheless a potentially interesting task, which might offer lessons on test development relevant to other researchers concerned with the CSE. It is often asserted that teaching and assessing Chinese learners of English neglects the speaking skill in order to concentrate on the testing of reading, writing and grammar. To the extent that this is true, then efforts to validate the assessment of the ability to speak English are laudable. As Luoma (2004: 1) points out, however, "Assessing speaking is challenging… because there are so many factors that influence our impression of how well someone can speak a language, and because we expect test scores to be accurate, just and appropriate for our purpose. This is a tall order." This paper illustrates some of the difficulties involved in assessing speaking.

Any attempt to assess speaking requires a clear understanding of the multi-faceted nature of speaking, as described in theories and models of speaking in a second or foreign language, which are operationalised in the constructs and sub-constructs that underlie that skill. It also involves developing speaking tasks that approximate to how and why people speak in any language. Assessing speaking necessitates the development of suitable test or assessment procedures. Scales need to be developed to allow or facilitate the assessment of a person's speaking skills. Developers also need to consider how performance on achievement and proficiency tests of speaking can be meaningfully reported to learners and their teachers, and research needs to explore how well the test results predict learners' performance in the real world. Readers are recommended to refer to Luoma's volume, especially Chapter 8 on "Ensuring a reliable and valid speaking assessment".

Exploring the adaptability of the CEFR in the construction of a writing ability scale for Test for English Majors

The paper by Zou and Zhang concerns Writing. The authors report on a case study to see whether and to what extent the CEFR is suitable for, or can be adapted to be suitable for, the construction of a writing ability scale for English majors in Chinese universities.

The two research questions asked have two sub-questions each, as follows:

1. Compared with descriptors from other sources, to what extent are CEFR descriptors adaptable in describing English majors' writing proficiency?

   (a) What can English majors do in writing?
   (b) Are CEFR illustrative writing descriptors suitable for profiling English majors?
2. How can we construct a writing ability scale of Can Do statements for English majors by utilising descriptors from various sources?
   (a) What are the procedures to be followed in developing a Can Do descriptor pool?
   (b) What criterion/criteria should be adopted to scale the descriptors?

In fact, one could argue that these are not really research questions, since potential answers are readily available in the CEFR literature. The real issue is whether such procedures can produce usable and meaningful results in the Chinese context. In other words, can the procedures used in developing the CEFR be replicated in the Chinese context?

The paper provides a description of a mixed-method approach to these issues in four stages: Stage 1 is the drafting of a questionnaire containing Can Do statements taken from the CEFR, A1 to C1, the Test of English Majors (TEM) syllabus and the Teaching syllabus, as well as the results of focus group discussions, as shown in detail in Appendix One. Stage 2 was the successful administration of the questionnaire to 194 teachers and PhD students and the creation of an initial descriptor pool. Stage 3 involved 35 teachers judging the level of difficulty of 40 descriptors with reference to 36 TEM student writing samples at pre-determined levels of proficiency, taken from four operational administrations of TEM. In Stage 4, small-scale interviews were conducted with eight university teachers for their opinions about the difficulty levels of the descriptors in order to enable the finalisation of “the writing ability scale profiling what students can do at different levels of proficiency”. Appendix One includes the mean values and the IRT results of the analysis, Appendix Two presents the logit values of the TEM writing scripts and Appendix Three presents the original questionnaire as administered.

This detailed presentation of the results is exemplary for its value not only in enabling the verification of the subsequent discussion of the results, but also and perhaps more importantly, in allowing replications of this study in other universities and contexts. Moreover, the detailed description of this study’s methodology will enable similar studies at lower levels of English education in China.

The final version of the writing scale, presented in Table 1, consists of three levels, with 10 descriptors at the highest level (Level 3), 13 descriptors at Level 2 and 14 descriptors at Level 1. Perhaps unsurprisingly, Level 3 descriptors broadly match the ability level required for senior students of English, Level 2 that of junior students of English and Level 1 are “mostly what freshmen majoring in English are supposed to be competent in”.

As with the other papers in this Special Issue, it remains to be seen whether further empirical evidence supports or disconfirms this division into three levels and whether the intended three levels are representative of the actual writing performance of English majors in China. It is conceivable that the heterogeneity of English education across such a varied country as China might complexify both the replications and the results. Nevertheless, this paper is a worthwhile beginning and basis for revision in future studies.

Calibrating the CEFR against the China Standards of English for college English vocabulary education in China

The paper by Zhao, Wang, Coniam and Xie explicitly takes the "CEFR as a reference point" and states the purpose of the study as "to conduct an external validation of the CSE vocabulary descriptors with reference to the CEFR vocabulary descriptors". They acknowledge at the start, however, that some of the CEFR "illustrative" descriptors have been criticised for opaqueness, for inconsistencies in describing vocabulary constructs, for lack of definition of terms and for the CEFR's language-neutral scope, which means that it has "little to say about the nature of vocabulary in particular languages, or about the nature of lexical ability" (Alderson, 2005: 192).

Moreover, the authors acknowledge that "a review of the CEFR and the College English Curriculum Requirements (CECR) (Ministry of Education, 2007) indicated that the vocabulary descriptors in both documents were inadequate to describe vocabulary knowledge for College English education in China".

Given these doubts, it is perhaps inappropriate to consider this paper as a validation, but rather as an interesting study which uses some of the CEFR vocabulary descriptors to explore whether experienced Chinese teachers of English at college level can agree on the levels at which the descriptors set in the CECR are suitable for scaling vocabulary in the CSE.

In fact, the results of the study showed considerable disagreement among the 22 teachers as to the levels of the descriptors (B1, B1+ and B2, or CSE5 to CSE8). One of the reasons for such disagreement might be that the participants were insufficiently familiar with the CEFR (or indeed the CECR) or that the particular descriptors chosen for the study were problematic. Indeed, 17 of the 39 descriptors chosen for the study proved to range over more than three levels (i.e. were rated either one level higher or one level lower than their original CEFR level—Table 6). Given that this was a pilot study to explore the feasibility of using the CEFR and CECR vocabulary descriptors to help develop the CSE scales, the authors conclude that "teacher judgment of the scales provides evidence of the CSE scales, and can be a source of valuable information for the future improvement of the CSE."

However, the authors recommend improving the study's methodology by using not only quantitative judgements and statistical analyses of the results, but also "qualitative follow-up interviews, to investigate in greater depth participants' perceptions in relation to their judgment making". They also recommend more extensive familiarisation with, and training on, the CEFR descriptors, and a wider sampling of the CEFR descriptors "since it is somewhat unlikely that a single one-off validation study will provide sufficient evidence of alignment (Martyniuk, 2010)". The sensible conclusion is reached that participant judgement alone is not sufficient for these purposes and that "multiple sources of evidence should also be provided to triangulate the empirical evidence so that consistent interpretation and modification can be provided".

Conclusion

This is a very ambitious project, with a time-scale of only 4 years from start to finish. This includes the actual scaling of the large datasets and empirical validation (intended to have been finished in 2016) and the publication of the first version of the CSE in 2017. In fact, this whole Special Issue of LTA is a report on a Work-in-Progress that will last for many years if not decades to follow up various angles on the data and to revise the first set of scales and develop them further. To give a comparison, the CEFR was launched in 2001 after two draft versions had been published in the mid-1990s. Related projects were underway both before that date and are still ongoing 16 years later. Recent projects include work on new descriptors for mediation (both translation and interpretation), as well as explorations of new methodologies for developing such frameworks and also replications of the original methodologies.

The conclusion reached in the paper on English vocabulary education applies to all of the empirical studies published in this Special Issue, or envisaged for the development of the CSE. Iterative cycles of testing and revision will be needed to develop more comprehensive illustrative descriptors which are representative enough adequately to reflect the range of CSE illustrative scales. These will then need to be subjected to empirical research, revision and further iterations of research.

One area that is occasionally mentioned in this collection, yet remains unexplored to date, is that of levels. Given the complexity of language proficiency and how it develops in learners over time, the effort to distinguish different levels of language proficiency is not without problems. One important question is what does it mean to be "at" a level, and how is a "level" to be defined? It is interesting to note that the paper on listening identified nine separate levels of listening ability (which, unsurprisingly, matched the number of levels in the proposed CSE), the paper on speaking test validation identified only six levels of speaking ability and the paper on the writing scales found only three levels of writing ability. If this seems to be a contradiction, then it needs to be remembered, for example, that the 54 students researched in the paper on speaking were first-year students doing compulsory courses in English, and therefore they represent a relatively homogeneous sample, compared with the large range of ability to speak English across the whole Chinese education system. It is highly likely that a relatively large number of levels or sub-levels of language proficiency will be needed to cover this range.

When considering how levels can be measured, it is worth noting that it was, in part, one of the effects of the CEFR that much research has looked at how to set standards (for which one might include levels) and how to develop cut scores. DIALANG (Alderson, 2005) was something of a pioneer in devising and implementing standard-setting procedures in the late 1990s in order to estimate whether a given learner is at A1, A2, B1, B2 or even the C levels according to the CEFR and the associated scales, Can Do statements and proficiency descriptors. Since then, experience in devising, running and analysing standard-setting procedures and the establishment of cut scores has grown rapidly, see, for example, the pioneering research by Papageorgiou (2007, 2014), the volumes by Csizek and Bunch (2007) and Csizek (2012) and the publications of The European Association of Language Testing and Assessment (EALTA) on standard-setting on its website www.ealta.eu.org and especially the Special Interest Group on the CEFR (http://www.ealta.eu.org). I believe that the CSE Project would benefit from further investigation of this literature.

This is an important project. I look forward to reading future reports of empirical research which will contribute to its success.

Acknowledgements

This Special Issue of the journal *Language Testing in Asia* reports on progress in the development of the China's Standards of English (CSE), with reference to the CEFR, but avoiding the limitations and perceived problems in the CEFR by localising the CEFR model to the realities, culture and education systems of China.

Author's contributions

The author read and approved the final manuscript.

Competing interests

The author declares that he has no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 6 October 2017 Accepted: 13 October 2017
Published online: 22 November 2017

References

Alderson, JC (2005). *Diagnosing foreign language proficiency: the Interface between learning and assessment*. London: Continuum See also www.dialangweb.lancs.ac.uk.

Csizek, GJ (Ed.) (2012). *Setting performance standards: foundations, methods and innovations*, (Second ed.,). Thousand Oaks: SAGE Publications.

Csizek, GJ, & Bunch, MB (2007). *Standard-setting: a guide to establishing and evaluating performance standards on tests*. Thousand Oaks: SAGE Publications.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly*, *1*(4), 253–266.

Luoma, S (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

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

Ministry of Education (2007). *College English curriculum requirements*. Beijing: Higher Education Press.

Papageorgiou, S. (2007). Setting standards in Europe: the judges' contribution to relating language examinations to the Common European Framework of Reference. Unpublished PhD thesis, Lancaster University.

Papageorgiou, S. (2014). Aligning frameworks of reference in language testing: the ACTFL proficiency guidelines and the Common European Framework of Reference for languages. *Language Testing*, *31*(2), 261–264.

Spöttl, C, Kremmel, B, Holzknecht, F, Alderson, JC. (2016). Evaluating the achievements and challenges in reforming a national language exam: the reform team's perspective. *Papers in Language Testing and Assessment*, *5*(1), 1–22.

Weir, CJ. (2005). Limitations of the common European framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 1–20.