BRITISH COUNCIL

TeachingEnglish

ELT Research Papers 16.01

Investigating the practice of The Common European Framework of Reference for Languages (CEFR) outside Europe: a case study on the assessment of writing in English in China

Ying Zheng, Yanyan Zhang and Youyang Yan

UNIVERSITY OF Southampton

WUHAN UNIVERSITY

ELT Research Papers 16.01

Investigating the practice of The Common European Framework of Reference for Languages (CEFR) outside Europe: a case study on the assessment of writing in English in China

Ying Zheng, Yanyan Zhang and Youyang Yan

ISBN 978-0-86355-806-1

10 Spring Gardens
London SW1A 2BN, UK

www.britishcouncil.org

Authors

Dr. Ying Zheng is a Lecturer at the Faculty of Humanities and Deputy Director of the Confucius Institute, University of Southampton. Before joining the University of Southampton in 2013, Ying worked as a Psychometrician and Director of Research (2009-2013) in the Language Testing Division of Pearson, London. Ying's area of work has covered language tests from China, Canada, and the UK. She has been serving as an external psychometric consultant over the years for international language testing organisations. She currently supervises MA/PhD students working in the fields of assessment literacy, large-scale testing and psychometrics.

Dr. Yanyan Zhang has been working at the English Department of Wuhan University since 2008. She teaches English and linguistic courses at both the undergraduate and postgraduate levels. Her major research interests include Second Language Acquisition, Applied Linguistics, Language Testing and Assessment. She has published two monographs and a dozen research articles in peer-reviewed journals in China and abroad. She has also made many presentations on her research at national and international conferences.

Yan Youyang is a post-graduate student in Wuhan University, China. He majored in English and got his Bachelor Degree in Wuhan University in 2013. He is now in the 3rd year of his MA programme in English Language & Literature. Youyang has learned several languages besides English, including German, French, and Japanese. His main research interests are language testing and English varieties.

Contents

1

Research background

Since the publication of the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR) in 2001, a large number of studies have been carried out to investigate the use and impact of the CEFR in various contexts. The CEFR, established and developed by the Council of Europe and its member states, describes language ability in a systematic and comprehensive way. It also provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks (Council of Europe, 2001). Ever since its publication, the CEFR has been playing an influential role in language education not only across Europe, but also worldwide.

The CEFR, which clearly classifies language users into several successive proficiency levels, allows language learners' progress to be measured at each stage of learning and on a life-long basis (Council of Europe, 2001). Buckland (2010) encourages the CEFR to be used for benchmarking learning outcomes. By aligning the internal levels of the Wall Street Institute International with the CEFR scales, he managed to provide the students and their potential employers with accurate benchmarks on students' learning outcomes. Researchers have also shown that the CEFR can be used to inform the learning and teaching of English, and classroom assessment. By mainly adopting a questionnaire survey and teacher interviews in Sweden, Oscarson and Oscarson (2010) found that the use of the CEFR in teaching could enhance both students' language awareness and their learning motivation, and could also help make assessments more transparent, as reported by the teachers under investigation.

As is described above, most of the previous research on the use of the CEFR focused on Europe. Despite the fact that the CEFR has been widely understood and applied in European countries, it is not so well-known outside Europe (Buckland, 2010). As a matter of fact, there is even less discussion on the application of the CEFR in Asian contexts. To our knowledge, little research has been done on the applications of the CEFR in the context of China. To address this gap, this study aims to explore the practicality of the CEFR, to examine whether it can be used to promote ELT in China at the tertiary level, and hopefully to align its standard to a common framework. To be specific, this study investigates the possible applications of the CEFR in terms of assessing the writing of English by Chinese university students.

2

Relevant literature

The CEFR: background

A multicultural and multilingual Europe calls for its citizens to communicate with each other across linguistic and cultural boundaries, while at the same time preserving and developing its cultural and linguistic diversity. Early in 1975, Trim put forward the ideal of 'free movement of men and ideas' and claimed that increasing and improving language learning is a major way to achieve this. He also advocated children's language learning in compulsory education. Such opinions were ambitious but revolutionary at that time. The idea that a promising Europe demands both more language learning and the learning of more languages lays the foundation of the CEFR, which associates plurilingualism with partial competences. (Little, 2006).

Plurilingualism, as mentioned above, refers to one's ability to speak more than one language. Partial competence is defined in two ways. Many people have learned two or more foreign languages. However, they may only have reached a level some way below complete mastery, though that level might be almost enough for daily communication. On the other hand, people may only develop a limited range of communicative skills, such as in listening and speaking only, or in reading only. The CEFR's promotion of plurilingualism as a central goal of language education policy reflects a significant development in the Council of Europe's thinking that corresponds to equally significant developments in Europe's linguistic situation, that is, making more languages available to learners, and recognising that different objectives may be appropriate for different learners and different languages.

The CEFR scales

The Common European Framework divides learners into three broad divisions: A, Basic user; B, Independent user; C, Proficient user. These three broad divisions can be further divided into six levels: *A1 Breakthrough, A2 Waystage, B1 Threshold, B2 Vantage, C1 Effective Operational Proficiency* and *C2 Mastery*. To be specific, *Breakthrough* indicates a basic ability to communicate and exchange information in a simple way. *Waystage* suggests an ability to deal with simple, straightforward information and begin to express oneself in familiar contexts. *Threshold* refers to the ability to express oneself in a limited way in familiar situations and to deal in a general way with non routine information. *Vantage* stands for the capacity to achieve most goals and express oneself on a range of topics. *Effective Operational Proficiency* focuses on the ability to communicate with the emphasis on how well it is done, in terms of appropriateness, sensitivity and the capacity to deal with unfamiliar topics. *Mastery* calls for the capacity to deal with material which is academic or cognitively demanding, and to use language to good effect at a level of performance which may in certain respects be more advanced than that of an average native speaker.

The six reference levels are becoming widely accepted as the European standard for grading an individual's language proficiency. The CEFR now has a major impact on language education worldwide. As Byram and Parmenter (2012) noted, the CEFR has become the most reliable reference for developing strategic language policy documents, which to a large extent prescribe curriculum planning and practical teaching materials.

Linking assessments to the CEFR

The challenge that education systems face is to ensure, as far as possible, that all modes of assessment work together to produce observations and judgements that are as accurate and comparable as possible. The CEFR aims to help describe the proficiency levels of existing standards, tests and examinations in order to facilitate comparisons between different systems of qualification (Council of Europe, 2001). Linking tests to international standards such as the CEFR is a way of establishing criterion-referenced validity, which is an essential concern in test development. The major international testing agencies have been quick in adopting its reference levels as a common metric, with evident gains in transparency and comparability (Little, 2011). In 2003, the Council of Europe published a manual for relating language examinations to the CEFR. Over the past decade, a large amount of work has been carried out on establishing the alignment of tests with the CEFR. Martyniuk's (2010) book gathered a series of studies that looked at linking a single test to the CEFR, as well as linking a suite of exams to the CEFR. The majority of these studies undertake the systematic stages of familiarisation, specification, standardisation, and empirical validation as recommended by the manual (Council of Europe, 2009). For example, Kantarcioglu et al (2010) reported on a study linking the Certificate of Proficiency in English (COPE) to the CEFR B2 level. O'Sullivan's (2010) study provided empirical evidence that aimed to confirm the link between a single test, the City & Guilds Communicator, and the CEFR B2 level. Downey and Kollias (2010) attempted to link the Hellenic American University's Advanced Level Certificate in English examination (ALCE) to the CEFR, with the results suggesting that the ALCE test is targeted at the C1 level of the CEFR.

The majority of the studies reviewed above are concerned with establishing validity evidence for the existing tests by linking to the relevant CEFR levels. Schilling (2004) claimed that, in addition to data-based statistical analysis of how a test can claim its alignment to the CEFR, test design decisions and the evidence that supports these decisions can also make a significant contribution to the establishment of validity. In other words, the descriptive apparatus that embodies the CEFR's action-oriented approach is intended to apply not only to the comparison of language examinations but to the design, development and modification of examinations and tests. Eckes et al (2005) provided a series of brief reports on the reform of language assessment in the Baltic States, France, Germany, Greece, Hungary, Poland and Slovenia. De Jong and Zheng (2015) report on how CEFR scales were operationalised in practice in the course of developing the Pearson Test of English Academic (PTE Academic), which entails the process of item writing, item seeding, rating scale development and human rater training.

The CEFR as a rating scale

Admittedly, the CEFR was not initially written to be considered as a test development standard. Except for the self-assessment grid, the scales in the CEFR were not originally developed for rating learners' performances. Most CEFR scales focus on defining elements such as tasks, activities and texts, which are typical features of scales designed for descriptive and reporting purposes. Thus, they lack typical features of scales designed specifically for assessment purposes such as concrete references to errors in learners' performance.

However, from 2009 to 2012, studies by Chen (2009), Kuiken et al. (2010), Carlsen (2010), Forsberg and Bartning (2010) and Eckes (2012) suggest it may be possible to use CEFR scales for rating purposes. Each of them applied a slightly different approach to CEFR-related rating, accumulating evidence of the practicability of the CEFR as a rating scale thanks to the coherence of the descriptors that form the CEFR levels. Harsch and Martin (2012) modified CEFR scales to make them more rater-friendly, also showing promising results in this regard.

CEFR scales for the writing of English

Powerful though the CEFR has been in both language learning and testing, its compatibility with findings from second language acquisition (SLA) research or its suitability for young learners remains uncertain and deserves attention (Little, 2007; North, 2007). Generally, rating scales are not commonly used in SLA studies (Tremblay and Garrison 2010). For example, can specific linguistic features be associated with specific proficiency levels? Such a question calls for empirical studies of the relationship between communicative L2 development, such as functions described in the proficiency levels of the CEFR and the development of the linguistic skills, like vocabulary and structures. For studies that examine links between linguistic features and the CEFR levels, a particularly important question is whether the straightforward approach of using CEFR scales to place learner performances on proficiency levels really works. In any case, the suitability for rating of any scale needs to be examined before the ratings obtained with it can be trusted. Thus, the validity of the CEFR scales for rating writing cannot be automatically assumed.

Not originally developed for rating learners' performances, the CEFR scales have however been used for that purpose. In the section above, we found that some previous studies showed promising results in this regard (e.g., Chen, 2009; Carlsen, 2010). Yet, those CEFR scales are more or less modified to be more feasible for raters, so as to ensure the quality of the rating. In 2014, Huhta et al. (2014) published a study which tended to validate the rating procedures in SLA research across two scales: an unmodified CEFR scale and a modification of the CEFR scale. The researchers in the study rated learners' performances on writing by using two scales, the CEFR scale and the Finnish National Core Curriculum for Basic Education (NCC). The CEFR scale was adopted without the wording being modified. The NCC scale was developed in Finland in the early 2000s with much of its content coming from the CEFR. It comprises 10 levels, by dividing the original CEFR scale levels into two or three sub-levels. The results showed that both the unmodified CEFR scale and the local modification of the CEFR scale were suitable for rating.

CET: background

The College English Test, better known as CET, is a national test of English as a foreign language in the People's Republic of China. The purpose of the CET is to examine the English proficiency of college students, including both undergraduate and postgraduate students in China, and ensure that students reach the required English levels specified in the National College English Teaching syllabuses. This test has existed in China for nearly 30 years and now has a huge test population of around 18 million people annually, which makes it the largest test of English as a foreign language in the world in terms of participants. It was held nationally twice a year in June and December. The CET consists of two levels, Band 4 (CET4) and Band 6 (CET6), which are the English levels that non-English major undergraduate and postgraduate students are supposed to reach respectively. The CET is mandatory for university students in China who are not English majors. Passing the CET is important for Chinese college students. It is also a prerequisite for a bachelor's degree in most universities. Many employers in China prefer applicants with a CET certification.

The CET was reformed in 2005 with the introduction of a new grading system from the previous maximum score of 100 points to a maximum score of 710. The lowest mark is a score of 290, on condition that the test-taker finishes all of the questions but gets them all wrong. According to the authority, the *passing point* was eliminated, and the qualification certificate was changed into a score report card, showing a more detailed report on each section, instead of an overall score. But conventionally, certificate-holders of 425 points, accounting for 60% of the total score, are considered to have passed the test. Only university students are allowed to take the test, instead of opening it to the general public.

The CET writing task and rating scale

In the writing section, students are asked to write a composition of no less than 120 words for the CET-4 and 150 words for the CET-6, based on the information provided (e.g., title of the topic, outline, situation, pictures, or graphs). The time limit is 30 minutes. It is worth noting that for the writing part of CET 4 and 6, the instruction has been changing in recent decades, from outline composition to proposition composition, and then to picture composition. The trend is to develop more diversified instructions.

Before 2011, almost all CET writing tasks were outline compositions. In outline compositions, the instruction gives students the title and detailed hints or materials. A trained student can easily write a three-paragraph essay following a certain format, introducing the topic in the first paragraph, giving detailed opinions in the second and concluding or summarising in the last.

Undoubtedly, this was a big step in the reform of CET writing tasks. However, much against the test writers' will, there was no obvious change in students' writing. Although some essays with excellent language proficiency and original thought stood out, most essays still adopted a certain format. Raters were still buried with sentences like 'Nowadays, the issue of the way to success has attracted a public concern. As to this matter, different people have different ideas', 'With the development of our society, more and more people want to get success. People's attitudes towards the way to success vary from person to person', and so on.

With the continuous effort to further prevent students from adopting a certain format or framework, the CET authority introduced picture composition in 2013. In picture compositions, a general topic and a cartoon are given in the instruction. Students are asked to describe the picture and explain its connotation first, and then give comments.

Not much research has been conducted into the rating scale for CET writing. Moreover, it is surprising to note that this rating scale has experienced little change since the first CET test in 1987, although the writing task has changed a lot. The scale divides students' writing into five levels, mainly according to the following four aspects: relevance to the topic, the expression of ideas, coherence, and the amount of language errors. The descriptor for each level is a general sentence, without further definition or explanation in detail. There is no proportion provided of the four aspects for raters to balance or focus on, so, in principle, the four aspects are considered of equal importance. Yet, since the other three aspects are more difficult for raters to control, most raters mostly rely on the amount of language errors, when judging the writing.

3

Research design

The research purposes of this study are threefold. Firstly, it intends to introduce the CEFR to Chinese ELT teachers and to familiarise them with the use of the CEFR in the context of China. Secondly, it attempts to explore the applications of the CEFR in China and provide Chinese ELT teachers with hands-on practice of using the CEFR to assess the written English of university students. Thirdly, this study hopes to evaluate the CEFR scales and investigate whether the CEFR scales are applicable to the Chinese context and whether any amendments are needed to the illustrative descriptors. Two research questions are addressed:

1. What is the current knowledge of the CEFR among Chinese ELT teachers? After training and practice, what are Chinese ELT teachers' perceptions of the CEFR?
2. To what extent do Chinese ELT teachers' ratings of English writing written by Chinese university students agree with the CEFR experts' ratings? How could the CEFR rating scales be applied to the assessment of Chinese university students' writing of English?

To achieve the above research aims, this study adopted a mix-method approach. Corresponding to the research questions, different methods were employed. The participants include around 40 ELT teachers, 120 students and 9 CEFR experts. The 40 ELT teachers are all randomly selected from Wuhan University in China. The 120 students are composed of 40 freshmen, 40 sophomores, and 40 juniors, all of whom are non-English majors from Wuhan University. The 9 CEFR experts come from academic institutions in the United Kingdom, namely the University of Southampton and the Pearson Education Group. The instruments include writings in English from the 120 students, the CET writing rating scale, the CEFR self-assessment grid for writing (Council of Europe, 2001), the CEFR general linguistic range for writing, a questionnaire given to the 40 teachers and an immediate interview to the 9 teachers. The CEFR general linguistic range for writing was specifically made for this project, based on the original scale.

Data collection procedure

Before rating

To investigate Chinese English language teachers' knowledge of the CEFR, this study adopts a questionnaire survey. About 40 ELT teachers from Wuhan University participated in the survey, in November 2014.

While rating

To examine how much agreement Chinese ELT teachers have with CEFR experts in terms of rating, writing data were collected and raters were invited to rate the writing samples. For this purpose, the study adopted a cross-sectional design. Three groups of students from Wuhan University were recruited to participate in the study, in early December 2014. The three groups of students consisted of 40 freshmen, 40 sophomores, and 40 juniors, with their English proficiency roughly at pre-CET 4, CET 4 and post-CET 4 levels, respectively. The students were asked to write an argumentative essay on a certain topic within half an hour. The test-takers are expected to write at least 300 words but no more than 350 words.

Writing Topic:

> *Technology and education (Online learning vs traditional education)*

Instruction:

> *Some people think that computers and the Internet are more important for a child's education than going to school. But others believe that schools and teachers are essential for children to learn effectively. Discuss both views and give your own opinion.*

Altogether 9 CEFR experts did the rating according to the CEFR writing criteria, using the analytic scales. The 9 CEFR experts were randomly and evenly divided into three groups, with 3 experts in each group responsible for rating 40 students' writing samples respectively.

In mid-December 2014, two CEFR experts from the UK were invited to train the 9 Chinese ELT teachers from Wuhan University on the CEFR and on how to rate the writing of English according to the CEFR scales. After the training, the 9 teachers also rated the students' writing. The 9 teachers were divided into three groups as well, with 3 people in each group rating 40 pieces of writing samples respectively. Unlike the CEFR experts group, the Chinese ELT teachers rated the writing samples on two different scales, first on the CET 4 writing scale, which was already familiar enough to them, and then on the CEFR writing scale. The rating on the CET 4 writing scale was finished before they attended the training, and the rating on the CEFR writing scale was carried out after the CEFR training in late December 2014. All of them managed to complete their rating within one week.

In total, each piece of writing was rated 9 times as a result, once by each of three CEFR experts and twice by each of three Chinese ELT teachers. When the rating was completed, scores from the CEFR experts and the Chinese ELT teachers were compared to see how the two groups of raters agree with each other. The comparison between the two sets of rating data, from CEFR experts and ELT teachers, is an experience of international collaboration. More importantly, through the training and rating exercise, ELT teachers in China can get a better understanding of the CEFR from CEFR experts and from their own hands-on practice.

Rating score comparison could be made not only horizontally, by comparing scores between CEFR experts and Chinese ELT teachers, but also vertically, by comparing scores across different English levels, namely, pre-CET 4, CET 4, and post-CET 4 levels. The purpose is to examine if there is any progression of agreement between CEFR experts and Chinese ELT teachers across levels.

In addition to the between-group comparison, within-group comparison is also undertaken. The comparison between the rating scores by the three CEFR experts in each group provides evidence on whether there is within-group consistency or agreement in the use of the CEFR writing scales by CEFR experts. Similarly, the comparison between the rating scores by the three Chinese ELT teachers in each group may demonstrate any agreement or disagreement among Chinese ELT teachers in their rating, according to the CET 4 and CEFR writing criteria. Moreover, such a comparison could also help establish a potential alignment between CET 4 and the CEFR writing scales.

After rating

To examine further how the CEFR rating scales can be applied to Chinese university students' writing assessment, a post-activity interview of teachers was given to the same nine Chinese ELT teachers in early January 2015, just a few days after they had finished their rating. They were invited to discuss their perceptions of the benefits of introducing the CEFR into China, and to reflect on the rating exercise they have completed. They were also encouraged to analyse the possible stumbling blocks preventing the CEFR from being applied. The interview was done on a one-to-one basis, and each interview lasted for about 20 minutes.

In brief, by answering the three research questions adopting the above methods, this study will help improve Chinese ELT teachers' understanding of the CEFR, familiarise them with the use of the CEFR for assessing writing in English, map the Chinese students' writing of English onto the CEFR and make possible amendments so that the CEFR concept that is currently dominantly used in Europe can be further extended to the largest ELT market in the world.

4

Results

Qualitative findings

The data from the post-activity interviews were analysed qualitatively. The following are the major findings from the questionnaire and post-rating interviews.

Current knowledge of the CEFR among Chinese ELT teachers

Unfortunately, the majority of English teachers in China know nothing about the Common Framework, and have never thought about the differences in grading standards in China and abroad, except for a few teachers who are aware of some of the differences between Chinese and European evaluation standards from their own experience. Before taking part in the project, most teachers had little idea about the CEFR. Some had never even heard of it. Teacher 1 says that in the past she didn't know the CEFR at all. Most of her fellow teachers didn't even know CEFR was such a grading system. T2 says, 'to tell the truth, I didn't know that before'. T5 says that he knew very little about this before. Possibly, teachers, who have little overseas experience, hardly have access to CEFR within China.

However, a few teachers had some experience in working as an examiner for the BEC, which made them aware of the difference between Chinese and European evaluation standards at first. T2 and T5 claim that they have been working as judges for the BEC speaking part, at both intermediate and advanced level. T2 says that she found BEC standards very similar to CEFR requirements, since both are from Britain. T5 says:

> *'It seems that there's a B1 grade in BEC. I'm curious about the relation between band B1 in CEFR and B1 in BEC.'*

Another teacher once taught Chinese in Europe, where she got in touch with a similar evaluation standard. She says,

> *'I taught Chinese in Ireland in 2002 and applied some rating criteria made in Europe, but unfortunately I am not sure whether it was CEFR.'*

Teachers' opinions of the training

Two CEFR experts from the UK were invited to train the nine Chinese ELT teachers from Wuhan University on the CEFR and on how to rate written English according to the CEFR scales. The teachers got to know the CEFR in more detail through the training. T1 says that it is almost the first time for her to get to know the CEFR. T7 says, 'The workshop held by experts from the University of Southampton last month made us get to know what CEFR is'. T8 says that she got a general impression. They all found the training very inspiring and showed great interest in CEFR. The training helped them understand the object, content, method, application and significance of CEFR. All believe that CEFR, as a valuable evaluation standard for us to refer to, helps to improve our grading system on students. T1 says that the training has broadened her eyes. T3 says:

> *'The trainers are fully prepared, so we trainees learned a lot.'*

T7 says:

> *'It's a great honour to be invited to join in this research project. We believe it's a good chance as well as a valuable experience… I think the training is helpful and constructive.'*

Teachers' opinions of the CEFR

Some teachers point out that since the CEFR is a potentially global evaluation standard for language, it must be not only more rational and scientific, but also set a shared goal for all language teachers and learners to strive for. The same student performs differently according to different evaluation standards, thus a unified, extensively acknowledged evaluation standard is crucial. In this way, language training can be more focused, and can reach the point more directly, instead of beating around the bush along the way. T5 says that since one person may get different scores under different scoring criteria, a widely-accepted standard allows us to be more target-focused. T3 says:

> *'I'm quite interested in how Europeans assess language competence of language learners, how the framework provides guidance for their teaching, and how to make joint efforts by sharing a criterion… CEFR must be useful. Europe is a huge place where language learning takes place all the time, especially English learning. English, as a useful tool for communication among European people, is widely used in almost every European country. They have rich and extensive practical experience since*

> *they use English every day despite their mother tongues. In contrast, there are only rare occasions where English is used in China. Thus, the framework developed by them must be more reasonable than that by our Chinese. I will bring it into my teaching regardless of different aims of exams.'*

Differences between the CET and the CEFR

The teachers believe that the CEFR and CET evaluation standards differ a lot in their nature, though they still share a few similarities. Generally speaking, the two standards have different aims and functions, and focus on different aspects of a student's ability. T3 says that CEFR and CET have different intentions. T9 says that their differences are big, for their focuses are different. T6 commented:

> *'At first, I intended to make a comparison between CEFR and CET. After the training, I found CEFR and CET quite different.'*

Form vs. content (linguistic competence vs. communicative competence)

First of all, CET mainly evaluates examinees on the language itself, while CEFR cares about one's ability to communicate and express. In other words, CET checks how much English a student has learned and mastered, while CEFR tends to find out how well you can use English to express yourself. In terms of one's language competence, CET focuses on quantity, while CEFR focuses on quality. T6 commented:

> *'CET measures the level of your English learning, and compares what you have learned with students all over the country. By contrast, CEFR grading standards emphasise the extent to which you can express yourself. CEFR and CET do not differ in their standards, but in their directions and their aims.... CET measures how much and how well you have learned, while CEFR evaluates whether you can put what you have learned into use. As a teacher, both systems are necessary in that I always need one test to check your learning as well as another to evaluate your ability to use the knowledge.'*

T3 even says:

> *'CET is only a test to check the effect of college English teaching. It doesn't really intend to figure out one's real English level. I don't think CET can be compared with CEFR.'*

Besides, CET emphasises form, but CEFR dwells more on content. Key words, or signal words, like 'on the one hand', 'on the other hand' may indicate good organisation in CET writing. Yet, CEFR welcomes coherent narration, true feelings and inspiring ideas, rather than talking nonsense in a certain format. In addition, CET focusses on detail, while CEFR focusses on integrity. CEFR cares about the practicality of the whole text. CEFR takes into account whether an essay fulfills its aim and function. But CET sometimes only examines the appropriateness of isolated sentences.

T1 says:

> *'CEFR requires examiners to take richness and complication of the content into consideration, while in CET; they only need to take language into account.'*

T8 says:

> *'Although CET also mentions structure and organisation in text, it actually refers to the organisation of language. CEFR examines whether, from the whole text and the overall structure, the meanings expressed and the central ideas have fulfilled the practical function.'*

Accuracy vs. logic

Last, but most importantly, evaluation relies heavily on language accuracy in CET, while the CEFR concerns the coherence in logic. Most teachers agree that language accuracy is the lowest level of evaluation. The CEFR encourages teachers to observe students' positive points instead of simply picking out grammatical mistakes. Students are encouraged to express themselves freely in different ways and by different methods, without being limited by so-called accuracy. T3 says:

> *'As for language errors, CET has a clear and specific standard covering all levels from 2 points to 14 points, such as too many serious errors, a large number of serious errors, some errors, or few errors. Yet, as far as I know, CEFR only mentions language errors in one certain level. In fact, CEFR doesn't put emphasis on accuracy. It calls for fulfillment of the communicative function and purpose in the writing task. The standards of CEFR have a broader coverage.'*

T5 says:

> *'The three frameworks, namely CET, BEC and CEFR, have different focuses. CET focuses on accuracy. BEC focuses on well-designed expressions and structure. Even if you make mistakes, you may be rewarded since you have the intention to achieve a higher level. CEFR focuses on communication and smooth expression. I think CEFR is more scientific.'*

T6 says:

> *'To be frank, I think it's more helpful to listening and speaking. For Chinese students, such a standard is especially crucial, for they pay too much attention*

> *to grammatical mistakes. But I feel that the biggest problem for Chinese students is their unidiomatic expression. Chinese students sometimes speak quite fluently, but no one knows what they are talking about. CET deliberately encourages students to use complicated expressions, even uncommon words. Your structure and diction are quite weird, which will never be used by a native speaker. Thus, under the CEFR framework, I shall lead my students to pay less attention to grammatical mistakes.'*

Perceptions of the CEFR

Strengths

We may observe the strengths of the CEFR from both micro and macro perspectives. From the micro perspective, it is reasonable and scientific, compared to other evaluation standards such as CET. We have already discussed this in the section above. Moreover, it is a widely-accepted standard which sets a shared goal for all language teachers and learners to strive for, as we mentioned above when reporting teachers' opinions on the CEFR.

From the macro perspective, we will discuss the influence on aspects other than learning, teaching and testing. As for the application of both evaluation standards, most teachers predict that CEFR would have a wider coverage, since it evaluates students' language competence from broader dimensions. T3 says that the strengths of CEFR definitely overweigh its weaknesses, for it aims higher than CET. Some teachers hold that CET only applies in college, but CEFR is useful abroad and in the job market.

T6 says:

> *'As far as I am concerned, in college, CET has got the advantage, since it's necessary to your learning progress. Yet, in job markets, or when going abroad, CEFR has more , it tests whether you are really able to interact with others in a company, or live abroad.'*

T6 explains later:

> *'Anyway, CEFR is introduced from abroad. Frankly speaking, CET is somewhat an isolated project. It's common that native speakers even don't understand what we regard as 'good' essays. For those 'bad' ones with low marks, native speakers feel quite okay if they understand well. In the training, we all give low marks on an essay, which actually gains 6 to 7 in IELTS. That essay, with mistakes in all person, number and tense, has not a single correct sentence. But it's understandable. So the communicative purpose is fulfilled from the point of view of native speakers. In this sense, the performance is successful! That's the reason why I'm interested in studying such a standard. We learn foreign languages to communicate with foreigners, to embrace the world. Only such an authoritative system abroad can really help us.'*

Weaknesses

Two weaknesses of CEFR are mentioned by interviewees. For one thing, the instructions in CEFR are not as easy as those in CET to understand and grasp. The descriptions in CEFR sound quite abstract, and seem to overlap a lot between neighbouring bands.

T6 says:

> *I think the description for each grade is not clear and precise enough. For example, you may give 6 and I may give 8, on the same CET essay. Then we'll finally reach an agreement, after discussing sentence by sentence, since the standard is specific enough. But I don't think this will be the case for CEFR. I believe, even through discussion, disagreements from teachers will still remain by CEFR standard.*

T7 says:

> *But the disadvantage lies in that CEFR criterion doesn't have the sample compositions. They just have the descriptor for each band. They only have the general linguistic range and writing band descriptor. But they have no sample writing for each band.*

For another, a few teachers doubt whether CEFR would be officially accepted in China. China is a place where government plays a dominant role everywhere. Therefore, without official recognition and authorisation, CEFR could hardly take effect.

T2 says:

> *'As far as I am concerned, when a testing system is put into practice, the authority and organisers are extremely important. I surely want Chinese tests to be internationalised. So teachers and students don't need to face so many different types of tests and questions. Only in this way will students be less burdened.'*

T4 says:

> *'I don't think it's very useful to teachers in public English department, for the university attaches more importance to CET. But it's more helpful to training centers. In China, we emphasize unification, such as unified teaching plans and syllabus. But it's much more flexible abroad. I don't think CEFR has a promising future here in China.'*

Quantitative findings

The survey data and rating data were analysed quantitatively. In answering the research question regarding ELT teachers' rating consistency in using both the CET-4 writing rating scale and the CEFR rating scale, several steps were taken. First, the CEFR scale was transformed into a numeric scale using the conversion in Table 1. Four plus levels were introduced to add granularity to the rating. Altogether, there are 9 levels transformed into a scale from 1 to 9:

Table 1: CEFR rating scale

CEFR	Numeric scale
C1+	9
C1	8
B2+	7
B2	6
B1+	5
B1	4
A2+	3
A2	2
A1	1

The 120 essays from three groups were evenly divided and rated by nine ELT teacher raters, therefore each essay was rated by three raters, using both the CET scale and CEFR scale (see Table 2 for the rater profile). The correlation of the average rating scores from both scales, as indicated in Table 3, is 0.90, with r-square =0.82, meaning that in rating these essays using the two rating scales, 82% of the variance can be explained by both scales. In comparing the rating consistency among the raters of the three groups, not much variation can be observed.

Table 2: ELT rater profile

	CET-1	CET-2	CET-3
Group 1 (Freshmen)	Rater 1	Rater 2	Rater 3
Group 2 (Sophomores)	Rater 4	Rater 5	Rater 6
Group 3 (Juniors)	Rater 7	Rater 8	Rater 9
	CEFR-1	**CEFR-2**	**CEFR-3**
Group 1 (Freshmen)	Rater 1	Rater 2	Rater 3
Group 2 (Sophomores)	Rater 4	Rater 5	Rater 6
Group 3 (Juniors)	Rater 7	Rater 8	Rater 9

Table 3: average rating scores from both scales

	CORR	R SQUARE
ALL STS	0.90	0.82
YEAR 1	0.94	0.89
YEAR 2	0.91	0.83
YEAR 3	0.93	0.86

However, discrepancies were identified by checking the rater consistency among the individuals. The highest correlation is 0.98, and the lowest is 0.73 (see Table 4).

Table 4: individual correlation

CEFR	CORR
Rater 1	0.95
Rater 2	0.95
Rater 3	0.73
Rater 4	0.98
Rater 5	0.91
Rater 6	0.98
Rater 7	0.75
Rater 8	0.74
Rater 9	0.75

Overall, the result of the intra-rater correlation is satisfactory, with the lowest correlation 0.73, over half of the rating variance being explicable by the two rating scales after a one-day workshop. Taking into account these raters' feedback on the differences they perceived to be existent between the two rating scales, introducing the CEFR scale seems to be a variable option in enriching teaching and evaluating practices in the Chinese ELT context.

One surprising result is that there seemed to be no progression of students' essay scores from year 1 students to year 3 students, on both the CET and CEFR rating scales (see Table 5). These results remained unexplained.

Table 5: Averages of the ratings

	CET	CEFR
YEAR 1	10.56	5.76
YEAR 2	9.61	5.26
YEAR 3	9.81	4.83

The same 120 essays were also rated by 9 CEFR experts. They are called ‘experts’ for several reasons. All 9 of them have varying degrees of familiarity with the CEFR. There were no standardisation activities carried out with this group of raters. They received the typed essay scripts online, together with the specific CEFR rating scale. They were reminded that they should refresh their knowledge of the CEFR manual, and it was pointed out that the scale that was used for this particular project involved several ‘plus levels’. Each of the 9 raters rated 40 scripts, and they returned their ratings to the principal researcher by email. The overall correlation between the ELT raters’ average ratings and CEFR experts’ average ratings was 0.52. In terms of the ratings of the three groups, the correlations were 0.57, 0.65, and 0.42 respectively.

Further investigation among the expert rating revealed that there was not much consistency among the expert ratings either, with an inter-rater correlation ranging from 0.23 to 0.71. The causes of this situation were probably because 1) no standardisation activities were carried out, to bring everyone on to the same page; 2) these 9 experts were from different backgrounds, carrying various degrees of familiarity with the CEFR scale; 3) the CEFR scale used for this project included *plus* levels in the rating activities, which added variability to this exercise.

5

Discussions and conclusions

Ever since the CEFR became an essential element of European language education, due to political backing to a larger extent, rather than users' active involvement (Chalhoub-Deville, 2009), it has been somewhat controversial. On the one hand, the CEFR has provided important shared concepts for discussing language use and learning. To be specific, the CEFR has promoted an action-oriented view of language, with criterion-referenced assessment based on proficiency levels, and the concept of language profiles. Moreover, it has also raised awareness of the principles of valid and fair assessment. However, it is not the case that applications of the CEFR are always necessarily valid. The CEFR has often been implemented in a normative fashion that violates its intended flexible, concertina-like use as a reference tool (North, 2007).

When the CEFR levels were being used for educational and political purposes, very often the CEFR scale was not appropriately applied. In practice, such as for standard setting or curriculum design, for high-stake purposes such as citizenship, or in policymaking, using the CEFR as a helpful tool, usually the process was neither conducted transparently enough nor based on adequate empirical evidence. The very generic nature of the CEFR has also been criticised by Galaczi (2013). Particularly, the scales have been criticised for ambiguities and inconsistencies, such as whether the descriptors can differentiate between different proficiency levels (Alderson, 2007).

In using the CET rating scale, most raters regard the rating process as an effort to find language errors, especially basic grammatical mistakes in subject/verb agreement, tense, plural forms, articles, collocation and so on. That might be one of the contributing reasons why Chinese students are considered making slow progress in developing their English writing skills. In order to reduce possible language errors, students tend to use 'the safest language', thus avoiding complexity and variety in diction and sentence structure. For instance, a student will not risk saying "A teaching career enjoys high popularity among college students", instead of simply saying "many college students want to be teachers". Moreover, with such a standard, students will not risk exceeding the required word length, for the more you write, the more possible mistakes you will make. However, 120 words are too few compared to international English tests, like TOEFL or IELTS, which require at least 250 words for an argumentation type of essay. Undoubtedly, we cannot expect fully developed ideas and rich supporting details in only 120 words. Thus, in CET writing, content, structure and logic may not be as highly valued in the rating process as they deserve to be.

In applying the CEFR scale to rating, following the Chinese ELT teachers' participation and our communication with them, the results showed that the CEFR had no popularity among Chinese ELT teachers, even with professors from such a top Chinese university. Both internal and external reasons may account for the present situation. Domestically, although the CEFR has been translated into Chinese and published in mainland China, it has gained no official recognition or authorisation. The CEFR plays no role in any official policy or activity in the education field. No official English tests in China relate to the CEFR. English textbooks written in China do not take the CEFR into account. There is certainly a tendency to introduce original English textbooks from abroad into Chinese basic education, but these textbooks are heavily adapted to the practical conditions in the Chinese context.

Globally, the teaching and testing of English possess a long history, and had grown mature long before the CEFR achieved wide-spread acceptance. The CEFR does not exert much influence on English teaching methodology, though it does promote some techniques such as task-based language teaching (TBLT). With China's gradual and continuous opening-up to the outside world, international English tests, like TOEFL or IELTS, have also become influential in China and have been attracting huge numbers of Chinese test takers. However, all these tests maintain their own evaluation and grading systems, and thus have no direct or obvious connection to the CEFR, although they are claimed to be linked to the CEFR.

However, we have to note that the above discussion does not mean that the CEFR has no popularity in China. Although the CEFR seems quite unfamiliar to Chinese ELT teachers in "official" education, it is still known to other language teachers elsewhere in China. As for English teachers in non-official institutions, such as training centres, the CEFR may sound much more familiar to them, since these non-official institutions may adopt quite different teaching and learning systems, such as using only original English textbooks from abroad. Similarly, we may find that the CEFR is also well-known to teachers of other European languages. Unlike the two leading international English tests, namely TOEFL or IELTS, international tests of other foreign languages such as TestDaf (German), TEF/TCF (French) and DELF/DALF (French) all adopt the CEFR framework, either in test levels or score results. Therefore, French and German teachers in China may know more about CEFR than English teachers. However, more research needs to be conducted in the above areas, to get a larger and clearer picture.

Since most of the participating teachers had minimal knowledge of the CEFR, the training turned out to be necessary and helpful. In the training, the teachers quickly understood the CEFR scale and realised its advantages. As very experienced ELT teachers, they knew clearly the deficiencies of domestic evaluation standards, so they soon found the CEFR to be a very good complement. The teachers also showed great interest and willingness to apply the CEFR to their English teaching. All in all, as far as Chinese ELT teachers are concerned, the CEFR is highly welcomed by them, and has promising prospects in China.

Besides finding the CEFR very useful, teachers all agree that the CEFR scale is totally different from domestic evaluation standards such as the CET rating scale. Generally speaking, it is commonly recognised that the CEFR is more reasonable and scientific than the CET rating scale. Several main differences are pointed out. First of all, the CET rating scale evaluates one's knowledge in English, whereas the CEFR highlights application and practice. This is probably because the CET was originally intended as a test to check the effectiveness of the teaching of English in a college. Secondly, according to the participating teachers of this study, the CET rating scale mainly focused its evaluating criteria on the language form level, while the CEFR not only takes into account language content, but also includes evaluating aspects that check on thought and logic development. The reason may be that the CET has a huge population of test takers, so the rater has to be quick and only has time to read on the surface. Finally, in CET, the score depends mostly on the amount of errors that a student makes in writing. The CEFR, in contrast, seems to have a lighter focus on errors.

Teachers also give suggestions for further improvement of the CEFR writing scale. The CEFR writing scale could seem a bit abstract and subjective. Sample essays with analysis should be provided for each level. To promote the application of the CEFR in China, a 'top-down' strategy should be adopted instead of a 'bottom-up' one. In China, teachers and students have no free choice but to follow the mainstream. They have little say on education policy, so they are not able to promote the application of the CEFR in China by themselves alone. Therefore, they urge the authorities in China to realise the advantages of the CEFR. More importantly, domestic English tests in China should be further reformed, to keep in line with international English tests. Only with official recognition and acceptance can the CEFR benefit ELT in China in the long run.

Several conclusions can be drawn, based on the rating data collected. First of all, the results demonstrate that there is potential in introducing the CEFR scale into the Chinese ELT context. With proper training, that familiarises the teachers with the CEFR scale, including its background and the rationale of the scale, and examples demonstrating the scale differences, ELT teachers showed satisfactory rating consistency using the two scales. Taking into account teachers' positive opinions on utilising this scale to complement what they already have, it is safe to conclude that getting the teachers more familiar with the CEFR scale could potentially bring positive impact to their teaching and evaluation practices.

A few caveats need to be pointed out, in carrying out similar types of research in the future. To be able to get much more comparable and reliable rating results, the issue of whether plus levels should be used is worth further investigation. While plus levels may enhance rating accuracy, it may also introduce variability among raters. It is also not clear why Year 1 students' scores are on average higher than Year 2 and 3 students. To be able to fully understand the usability of the CEFR scale among students of different abilities, data from a larger sample of students' essays is needed. In addition, it would be advisable to carry out standardisation activities among all raters involved.

The preliminary results of this study are meaningful in the sense that it explored an approach to transforming CET scores to the CEFR scale. With more data collected and results confirmed, a transformation table could be created to enable CET-4 stakeholders to interpret scores in relation to the CEFR can-do statements and relevant descriptors. Future studies can expand the application of the CEFR scale to the rating of speaking, to explore whether by emphasising communicative competence, the CEFR scale can be used in the context of assessing speaking in China.

References

Alderson, JC (2007) The CEFR and the need for more research. *Modern Language Journal* 91, 659-663.

Buckland, S (2010) Using the CEFR to benchmark learning outcomes: a case study, in Mader, J and Urkun, Z (eds) (2010) *Putting the CEFR to Good Use* (4-10). University of Kent.

Byram, M and Parmenter, L (2012) *The common European framework of reference: The globalisation of language education policy.* Multilingual Matters, Bristol.

Carlsen, C (2010) Discourse connectives across CEFR-levels: A corpus based study, in Bartning, I, Martin, M, and Vedder, I (eds) *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (191-209). EUROSLA Monograph Series, 1.

Chalhoub-Deville, M (2009) Content validity considerations in language testing contexts, in R. W. Lissitz (ed) *The concept of validity: Revisions, new directions and applications.* Charlotte, NC: Information Age Publishing.

Chen, YH (2009) *Investigating lexical bundles across learner writing development.* (Doctoral dissertation). Lancaster University, UK.

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 Languages: Learning, Teaching, Assessment.* CEFR: A Manual. Language Policy Division, Strasbourg.

De Jong, J and Zheng (2015) Linking to the CEFR: Validation using a priori and a posteriori evidence, in Banerjee J and Tsagari D (eds) *Contemporary Second Language Assessment.* London/New York Continuum.

Downey, N and Kollias, C (2010) Mapping the Advanced Level Certificate in English (ALCE) examination onto the CEFR, in Martyniuk W (ed) *Aligning Tests with the CEFR: reflections on using the Council of Europe's draft Manual.* Cambridge University Press. Cambridge.

Eckes, T, Ellis, M, Kalnberzina, V, Zorn, K, Springer, C, Szollás, K, et al. (2005) Progress and problems in reforming public language examinations in Europe: cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing,* 22(3), 355-377.

Eckes, T (2012) Examinee-centered standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling,* 54(3), 257-283.

Forsberg, F and Bartning, I (2010) 'Can linguistic features discriminate between the communicative CEFR-levels? A pilot study of written L2 French', in Bartning, I, Martin, M and Vedder, I (eds) *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (133-157). EUROSLA Monograph Series, 1.

Galaczi, E (2013) Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics,* 1-23. doi:10.1093/applin/amt017

Harsch, C, and Martin, G (2012) Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing,* 17, 228-250.

Huhta A, Alanen R, Tarnanen M, Martin M and Hirvelä T (2014) Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing* July 2014 31: 307-328, first published on April 11, 2014 doi:10.1177/0265532214526176.

Kantarciouglu, E, Thomos, C, O'Dwyer, J and O'Sullivan, B (2010) 'Benchmarking a high-stakes proficiency exam: the COPE linking project', in W. Martyniuk (ed) *Aligning Tests with the CEFR: reflections on using the Council of Europe's draft Manual.* Cambridge University Press. Cambridge.

Kuiken, F, Vedder, I, and Gilabert, R (2010) 'Communicative adequacy and linguistic complexity in L2 writing', in Bartning I, Martin M and Vedder I (eds), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (81-99). EUROSLA Monograph series 1.

Little, D (2006) The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching* 39, 167-190.

Little, D (2007) The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal,* 91, 645-655.

Little, D (2011) The Common European Framework of Reference for Languages: A research agenda. *Language Teaching,* 44, 381-393.

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

North, B (2007) The CEFR illustrative descriptor scales. *Modern Language Journal,* 91, 656-659.

O'Sullivan, B (2010) The City & Guilds Communicator examination linking project: a brief overview with reflections on the process, in Martyniuk W (ed) *Aligning Tests with the CEFR: reflections on using the Council of Europe's draft Manual.*

Oscarson, AD and Oscarson, M (2010) 'Using the CEFR in the foreign language classroom', in Mader, J and Urkun, Z (eds) *Putting the CEFR to Good Use* (83-91). University of Kent.

Schilling, SG (2004) *Conceptualizing the validity argument: An alternative approach.* Measurement, 2, 178-182.

Tremblay, A and Garrison, MD (2010) 'Cloze tests: A tool for proficiency assessment in research on L2 French', in Prior MT, Watanabe Y and Lee S-K. Lee (eds) *Selected proceedings of the 2008 Second Language Research Forum. Exploring SLA perspectives, positions, and practices* (73-88). Somerville, MA: Cascadilla Proceedings Project.

Trim, JLM (1975) *Foreword,* in van Ek JA, The Threshold Level. Strasbourg: Council of Europe, i-iii.

This document is an output from the ELT Research Award scheme funded by the British Council to promote innovation in English language teaching research. The views expressed are not necessarily those of the British Council.

www.teachingenglish.org.uk/publications

ISBN 978-0-86355-806-1