Language Testing in Asia

RESEARCH Open Access

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

Wen Zhao[1*], Boran Wang[1], David Coniam[2] and Bingxue Xie[1]

* Correspondence:
dianawen@hotmail.com;
zhaowen@fsc.neu.edu.cn
[1]Northeastern University, Shenyang, China
Full list of author information is available at the end of the article

Abstract

Background: The Common European Framework of Reference (CEFR) (Council of Europe 2001) has, over the past decade, come to be widely used as a reference tool for teaching, learning and assessment (Alderson 2002; North 2014). The focus of the current study is on scaling the China Standards of English (CSE) vocabulary descriptors for College English education in Mainland China, where College English education refers to English language education for non-English major students at tertiary level. 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.

Method: On the basis of the CEFR and CECR descriptors, a pool of 39 descriptors were collected and categorized. Twenty-two English teachers from a Mainland China university were invited to participate in the study. They were first given the CEFR and CECR vocabulary descriptors, after which they were asked to scale the descriptors with the CEFR as the reference point. Multi-Faceted Rasch Measurement (MFRM) was used to validate teachers' scaling of the vocabulary descriptors.

Results: The MFRM analysis showed that while the descriptors at C1 level were generally ranked as expected, teachers had difficulties in ranking descriptors at the CEFR B1, B1+ and B2 levels.

Conclusion: The study indicates that teacher judgement of the scales provides evidence of the CSE scales, and can be a source of valuable information for the future improvement of the CSE.

Keywords: CEFR, Vocabulary descriptors, College English, China Standards of English, Validation

Background

The CEFR provides "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe 2001, p. 1). The influence of the CEFR has been "widespread and deep, impacting on curricula, syllabuses, teaching materials, tests and assessment systems and the development of scales of language proficiency geared to the six main levels of the CEFR" (Alderson 2002, p. 8). Its impact in Asia can be seen in such countries and regions as Japan (Negishi, Takada, and Tono 2013), Korea (Finch 2009), and Taiwan (Wu 2012).

As elsewhere around the world, English language teaching, learning and assessment in Mainland China are undergoing substantive change towards the establishment of a common framework of English language ability scales - the China Standards of English (CSE). This change is in answer to the pressing needs of constructing a transparent and coherent framework, taking the CEFR as a reference point, since a wide range of language curricula and assessments currently exist at different stages and levels of education (e.g., Han 2006; Yang and Gui 2007). For non-English major students' College English education in China, the College English Curriculum Requirements (CECR) (Ministry of Education 2007) serve as a guideline for College English teaching, learning and assessment. College English education in Mainland China refers to English language education for non-English major students at tertiary level. As a response to educational and curriculum changes, a trial version of the CECR was first issued in 2004 by the Ministry of Education in Mainland China. Following a 3-year trial of experiment and revision, a revised version was launched in 2007. In addition to its guiding role for College English teaching, learning and textbook writing, the CECR also serves as a guideline for the nation-wide CET (College English Test) Band 4 and Band 6 tests.

Along with the CEFR, the aim of the CECR is to cultivate learners' communicative language competence, with vocabulary knowledge considered key to language comprehension and communicative ability (Stæhr 2008). In developing a new set of vocabulary descriptive scales for the CSE, in particular CSE vocabulary descriptors for College English in China, an outline of the descriptive vocabulary scales of the CEFR and the CECR will first be described to lay out the background to the study.

Review of vocabulary descriptors

Vocabulary descriptor scales in the CEFR

The core conceptual framework of the CEFR consists of a taxonomic descriptive scheme and Common Reference Levels. The *descriptive scheme* covers domains of language use, communicative language activities, strategies, and communicative language competences for analyzing what is involved in language use and language learning. The *Common Reference Levels* describes proficiency in terms of three broad levels of basic user (A1 = Breakthrough; A2 = Waystage), independent user (B1 = Threshold; B2 = Vantage), and proficient user (C1 = Effective Operational Proficiency; C2 = Mastery) and in scales of illustrative descriptors across five qualitative categories: *Range*, *Accuracy*, *Fluency*, *Interaction* and *Coherence* (North 2014). These scales are of practical value in assessing learning and achievement (Alderson 2004).

Among the CEFR's 53 illustrative scaled descriptors, there are two qualitative categories: *Range* and *Control*, which are used to describe a learner's vocabulary knowledge (see Table 1, adapted from the Council of Europe 2001, p. 112).

Table 1 lists vocabulary descriptor scales of *range* and *control* at three CEFR bands and six levels. The descriptors used in the scales have all been empirically validated in terms of teachers' perceptions of how different levels of actual learner performance might be most consistently described. Each descriptor is stated in positive terms, and presents an independent criterion. The table indicates that in the CEFR learners are

Table 1 Vocabulary descriptors in the CEFR

Bands	Levels	Range	Control
Proficient User	C2	Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.	Consistently correct and appropriate use of vocabulary.
	C1	Has a good command of a very broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.	Occasional minor slips, but no significant vocabulary errors.
Independent User	B2	Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.	Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.
	B1	Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events.	Shows a good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.
Basic User	A2	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.	Can control narrow repertoire dealing with concrete everyday needs.
	A1	Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.	No descriptor available.

expected to *know*, *recognise* and *produce* a broad lexical repertoire to complete various tasks in different domains, contexts and topics.

The *range* of vocabulary expands as levels of proficiency advance. The lexical repertoire at the A levels covers words, phrases and everyday expressions, which are mostly an indicator of the breadth of vocabulary knowledge. The repertoire at the B levels describes a much wider range of vocabulary connected to fields and most general topics. The repertoire at the C levels, moreover, covers more idiomatic expressions, colloquialisms and connotative meanings, which focus more on the depth of vocabulary knowledge. The *control* of vocabulary states the degree and extent of vocabulary mastery across levels, with no description of vocabulary control contained at the A1 level.

The CEFR, however, is a general reference document which is at times somewhat problematic to use due to its language-neutral scope, and due to the fact that it "has little to say about the nature of vocabulary in particular languages, or about the nature of lexical ability" (Alderson 2005, p. 192). It is "a concertina-like reference tool, not an instrument to be applied" (North 2007). Although the CEFR descriptors state what learners at a certain level can do, very little is stated about what they should *know* in order to carry out related language tasks.

The CEFR, as a general reference document, has been criticized for the opaqueness of some descriptors, inconsistencies in its terminology use (Alderson *et al.* 2006), and insufficiency in describing vocabulary constructs (Huhta and Figueras 2004). Many terms in the CEFR are undefined, and there are problems with the wording of some descriptors. For instance, it is not easy to decide what is 'simple' and what is 'very simple'. Similar statements in the descriptors are found at different levels or some terms appear only at certain levels (Alderson, Kuijper, Nold, and Tardieu 2006, p. 9). Weir (2005, p. 12) observes that "the CEFR provides little assistance in identifying the breadth and depth of productive or receptive lexis that might be needed to operate at the various levels." The CEFR, moreover, is more a user-oriented set of scales than a constructor-oriented set of scales (Alderson 1991). Even the developers of the CEFR admit that its scales are primarily a taxonomy that is aimed at, and makes sense to, practitioners (North and Schneider 1998, p. 242–243).

Vocabulary descriptor scales in the CECR

Like the CEFR, the CECR adopts a functional approach in describing language skills and linguistic knowledge. The descriptive scheme in the CECR consists of three levels of requirements: *Basic*, *Intermediate* and *Advanced*. The vocabulary knowledge covers two dimensions: *Range* and *Control* (see Table 2) (adapted from Ministry of Education 2007, p. 3–5).

Table 2 indicates that, in College English education in China, learners are expected to *have a command* and *make skillful use* of a broad lexical repertoire to make meaningful communication. The *range* of vocabulary extends as levels of proficiency progress, as shown in its reference wordlist, phrase list and wordlist of active words. The CECR reference wordlist lays out its set of lexical items at three levels, with items at the *Basic Requirement* level unmarked, items at the *Intermediate* and *Advanced Requirement* levels marked with different symbols respectively (*for the Intermediate Requirement level; Δ for the Advanced Requirement level). The phrase list includes phrases and verbal phrases, with no inclusion of idioms, collocations or word chunks. The wordlist of active words is mainly a brief list of high frequency words.

The vocabulary repertoire at the *Basic Requirement* level covers 4,795 words, 700 phrases and expressions, the repertoire at the *Intermediate Requirement* level a much larger size of vocabulary (6,395 words, and 1,200 phrases and expressions), and the repertoire at the *Advanced Requirement* level a moderate increase of nearly 7,675 words and 1,870 phrases and expressions. The *range* of three levels of requirements gives only

Table 2 Vocabulary descriptors in the CECR

Levels of requirements	Range & control
Basic	Has a command of a lexical repertoire of 4,795 words and 700 phrases (including the vocabulary learned at senior secondary education), among which 2,000 words are active vocabulary, which a learner should be able to make skillful use in spoken and written English on the basis of recognition.
Intermediate	Has a command of a lexical repertoire of 6,395 words and 1,200 phrases (including the vocabulary learned at senior secondary education and vocabulary learned at Basic Requirement level), among which 2,200 words are active vocabulary (including the vocabulary learned at Basic Requirement level).
Advanced	Has a command of a lexical repertoire of 7,675 words and 1,870 phrases (including the vocabulary learned at senior secondary education, vocabulary learned at Basic and Intermediate Requirement levels), among which 2,360 words are active vocabulary (including the vocabulary learned at Basic and Intermediate Requirement levels).

quantitative descriptions of vocabulary size, with the inclusion of the vocabulary learned at senior secondary education and vocabulary learned at previous levels. The depth of vocabulary knowledge and tasks such as collocation, semantic meaning, and word formation are not stated in the document. Other than the description of vocabulary size, the domains, situations and topics relevant to vocabulary use are not stated in the CECR wordlist.

The CECR is organized by dictionary headword on the basis of corpus-based frequency information, with reference to the Bank of English (COBUILD Corpus). The CECR wordlists contain only lexical items with no provision of phonetic pronunciation, grammar and usage information, word definitions, dictionary examples and corpus-based learner examples. The CECR phrase list is also arranged alphabetically, with no word senses provided. It is therefore incumbent upon material writers, test developers as well as teachers to determine at what level and in what sense different lexical items should be selected or included.

The *control* of vocabulary, however, is only stated at the *Basic Requirement* level, requiring learners to be "able to make skillful use in spoken and written English". At the *Intermediate Requirement* and *Advanced Requirement* levels, there are no illustrative descriptors pertaining to vocabulary control. In comparison with the descriptors in the CEFR, the descriptors in the CECR are even more inadequate, inconsistent and underdefined in describing the constructs of vocabulary knowledge.

The analyses of the vocabulary descriptor scales in both the CEFR and the CECR indicate that the descriptors in both documents are for general rather than specific purposes. Moreover, these descriptors are not sufficient, and other descriptors need to be taken into account in developing CSE vocabulary descriptors for College English education in China.

Method

The purpose of the study is to conduct an external validation of the CSE vocabulary descriptors with reference to the CEFR vocabulary descriptors. The development of the CSE vocabulary descriptors will hence provide a more transparent, coherent and consistent guideline for College English teaching, learning and assessment in China, enrich the linking practice currently practiced in the development of the CSE, and make scales of vocabulary knowledge and their descriptors comparable by using the CEFR as the reference point. To this end, the research question in the current study may be framed as:

How well do the CECR vocabulary descriptors align with those of the CEFR?

Participants

In mid 2016, 22 English teachers from a mainland China university were invited to take part in the study. All were experienced female English teachers who had knowledge of College English teaching. Five of them had taught College English between 5 and 9 years, seven between 10 and 19 years, eight between 20 and 29 years, and two between 30 and 39 years. In terms of qualifications, 21 held an MA, and one a BA. All were familiar with the CECR, CET (College English Test) Band 4 and Band 6.

Procedure

Given the critical review of the inherent weaknesses of the vocabulary descriptive scales in the CEFR and the CECR, a pool of illustrative scales of vocabulary descriptors was collected, with the intention of covering three levels of requirements (i.e., Basic Requirement, Intermediate Requirement, and Advanced Requirement). These three levels roughly correspond to three CEFR levels (i.e., B1+, B2 and C1) and are equivalent to CSE's 6–8 levels:

- CSE1 (CEFR A1) is specified as the target for the end of primary school,
- CSE3 (CEFR A2) for the end of junior secondary school,
- CSE5 (CEFR B1) for the end of senior secondary school,
- CSE6 (CEFR B1+) for the Basic Requirement level of the CECR,
- CSE7 (CEFR B2) for the Intermediate Requirement level of the CECR, and
- CSE8 (CEFR C1) for the Advanced Requirement of the CECR.

The scope of the current study is limited to CSE5, CSE6, CSE7 and CSE8, as CSE5 is the entry English level for tertiary education. The scale construction followed the methods adopted in the scale construction of the CEFR. A pool of existing scaled descriptors were collected, whose sources are DIALANG (Alderson 2005), the CEFR in Finnish AMMKIA (North 2014, p. 79–80), and the CECR. DIALANG, a project which was explicitly developed using the CEFR as the basis (Alderson 2005), provides online diagnostic tests of listening, reading, writing, structures, and vocabulary in 14 languages at six CEFR levels. DIALANG vocabulary incorporates four dimensions of word meaning (i.e., word meaning, semantic relations, combinations, and word formation) in creating tasks. The AMMKIA descriptors for vocabulary were also scaled to the CEFR levels with considerable detail and the inclusion of vocabulary size (Kaftandjieva and Takala 2002). The vocabulary descriptors in both DIALANG and the Finnish AMMKIA were developed with the CEFR as the basis and contain more detailed illustrative descriptors.

In terms of *vocabulary size*, both the DIALANG and AMMKIA projects include vocabulary sizes at different levels. The suggested vocabulary size for C1 is 5,000 words. The vocabulary for B2 is, however, only 2,500–3,000 words, which might appear somewhat limited in number in comparison with the CECR and the English Curriculum of Senior Secondary Education (ECSSE) (Ministry of Education 2006). According to the ECSSE and the CECR, the vocabulary ranges for the suggested four levels are: 3,000 words for CSE5, 5,000 words for CSE6 (Basic Requirement level in the CECR), 6,000 words for CSE7 (Intermediate Requirement level in the CECR), and 8,000 words for CSE8 (Advanced Requirement level in the CECR). According to Hirsh and Nation (1992) there are two thresholds of vocabulary use: 2,000 word families is sufficient for 95% of typical texts encountered; 5,000 word families is sufficient for 99% of typical texts and for 'pleasurable reading'. In both DIALANG and the Finnish AMMKIA, 5,000 words is associated with C1, and 6,000 words with C2 (North 2014). In the CECR, however, 5,000 words is the minimum requirement for the Basic Requirement Level or CSE6. In the Chinese context, the vocabulary size issue was thus adapted according to the CECR to be in harmony with the teaching, learning, textbook writing and assessment. In addition to vocabulary breadth or vocabulary size, vocabulary depth is also taken into account. Verbs such as *know*, *recognize,* and *produce* are used to describe what learners are expected to do at different levels of proficiency (Alderson 2002).

The descriptors provide both quantitative and qualitative information. The collected vocabulary descriptors were further moderated to form a bank of 39 positively-worded descriptors, among which 8 descriptors were from B1 (SCE5), 7 from B1+ (SCE6), 11 from B2 (SCE7), and 13 from C1 (SCE8) (see Appendix 1).

Data collection

The 22 participants involved in the study were first given the CEFR general descriptive scales as well as the CEFR vocabulary descriptive scales. They were also given the CECR descriptive scales and its vocabulary descriptive scales. They were then given a brief introduction to the CEFR and the CECR to provide them with a clear overview of the descriptive scales.

A questionnaire was then prepared on the basis of the pool of 39 descriptors. The descriptors, containing both quantitative and qualitative information, were ordered according to the degree of cognition and degree of difficulty. The qualitative descriptors were sequenced before the quantitative descriptors on a four-point Likert scale, representing the four CEFR levels and the corresponding CSE levels (i.e., 1 = B1-CSE5, 2 = B1 + -CSE6, 3 = B2-CSE7, and 4 = C1-CSE8).

Participants were then asked to complete the survey online, and to scale the descriptors to appropriate levels on the four-point scale. The survey ensured complete anonymity, with all information used solely for purposes of the current study. No other person was permitted access to the information.

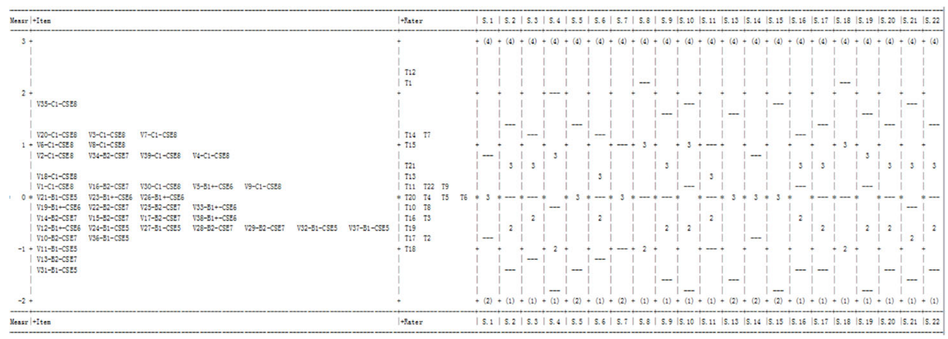
Data analysis

Linking was carried out via Multi-Faceted Rasch Measurement (MFRM) (Linacre and Wright 1994). MFRM is based on Item Response Theory (IRT), a branch of Latent Trait Theory. The advantage of MFRM lies in the fact that all facets (i.e., items [descriptors in the current instance], persons [participating teachers in the current instance] and judge [the participants' use of the rating scales]) can be compared on a common linear 'logit' scale (McNamara 1996). MFRM has been used to obtain information about severity and consistency of participants, the use of rating scales and items, as well as in studies investigating scales (North 2000).

The software FACETS 3.67 (Linacre 2005) produces a graph known as the 'all-facet vertical ruler map' or 'all-facet vertical summary'. FACETS also produces a measurement report for each facet of measurement. Fit statistics indicate how well the empirical data fit the measurement model's requirements. When fit values suggest that the data fit the Rasch model to an acceptable extent, unidimensionality is upheld. The fit values of the descriptors in the FACETS determined which descriptors would subsequently be included in the illustrative scales. Logit values were also considered in setting cut scores between adjacent levels (Papageorgiou 2009). The main output from the FACETS program is in the form of an all-facet overview and the facet reports, presented through charts and tables.

Results

The FACET ruler map is a useful tool in that it summarizes the position of the elements of each facet on the logit scale. It presents the information of the scattering of descriptors, participant ability and their use of rating scales (see Fig. 1).



```
-----+--------------------------------------------------------------------------------------------+---------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Measr|+Item                                                                                       |+Rater               | S.1 | S.2 | S.3 | S.4 | S.5 | S.6 | S.7 | S.8 | S.9 |S.10 |S.11 |S.13 |S.14 |S.15 |S.16 |S.17 |S.18 |S.19 |S.20 |S.21 |S.22 |
-----+--------------------------------------------------------------------------------------------+---------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
    3+                                                                                            +                     + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) +
     |                                                                                            |                     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
     |                                                                                            |                     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
     |                                                                                            | T12                 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
     |                                                                                            | T1                  |     |     |     |     |     |     |     | --- |     |     |     |     |     |     |     |     | --- |     |     |     |     |
    2+                                                                                            +                     +     +     +     + --- +     +     +     +     +     +     +     +     +     +     +     +     +     +     +     +     +     +
     | V35-C1-CSE8                                                                                |                     |     |     |     |     |     |     |     |     |     | --- |     |     |     | --- |     |     |     |     |     | --- |     |
     |                                                                                            |                     |     |     |     |     |     |     |     |     | --- |     |     | --- |     |     |     |     |     | --- |     |     |     |
     |                                                                                            |                     |     | --- |     |     | --- |     |     |     |     |     |     |     |     |     |     | --- |     |     | --- |     | --- |
     | V20-C1-CSE8  V3-C1-CSE8   V7-C1-CSE8                                                       | T14  T7             |     |     | --- |     |     | --- |     |     |     |     |     |     |     |     | --- |     |     |     |     |     |     |
    1+ V6-C1-CSE8   V8-C1-CSE8                                                                    + T15                 +     +     +     +     +     +     + --- +  3  +     +  3  + --- +     +     +     +     +     +  3  +     +     +     +     +
     | V2-C1-CSE8   V34-B2-CSE7  V39-C1-CSE8  V4-C1-CSE8                                          |                     | --- |     |     |  3  |     |     |     |     |     |     |     |     | --- |     |     |     |     |  3  |     |     |     |
     |                                                                                            | T21                 |     |  3  |  3  |     |     |     |     |     |  3  |     |     |     |     |     |  3  |  3  |     |     |  3  |  3  |  3  |
     | V18-C1-CSE8                                                                                | T13                 |     |     |     |     |     |  3  |     |     |     |     |  3  |     |     |     |     |     |     |     |     |     |     |
     | V1-C1-CSE8   V16-B2-CSE7  V30-C1-CSE8  V5-B1+-CSE6  V9-C1-CSE8                             | T11  T22  T9        |     |     |     |     |     |     |     |     |     | --- |     |     |     |     | --- |     |     | --- |     |     |     |
    0* V21-B1-CSE5  V23-B1+-CSE6 V26-B1+-CSE6                                                     * T20  T4   T5   T6   *  3  * --- * --- *     *  3  * --- *  3  * --- * --- *     * --- *  3  *  3  *  3  *     * --- * --- *     * --- *     * --- *
     | V19-B1+-CSE6 V22-B2-CSE7  V25-B2-CSE7  V33-B1+-CSE6                                        | T10  T8             |     |     |     | --- |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | --- |     |
     | V14-B2-CSE7  V15-B2-CSE7  V17-B2-CSE7  V38-B1+-CSE6                                        | T16  T3             |     |     |  2  |     |     |  2  |     |     |     |     |  2  |     |     |     |  2  |     |     |     |     |     |     |
     | V12-B1+-CSE6 V24-B1-CSE5  V27-B1-CSE5  V28-B2-CSE7  V29-B2-CSE7  V32-B1-CSE5  V37-B1-CSE5  | T19                 |     |  2  |     |     |     |     |     |     |  2  |  2  |     |     |     |     |     |  2  |     |  2  |  2  |     |  2  |
     | V10-B2-CSE7  V36-B1-CSE5                                                                   | T17  T2             | --- |     |     |     |     |     |     |     |     |     |     |     | --- |     |     |     |     |     |     |  2  |     |
   -1+ V11-B1-CSE5                                                                                + T18                 +     +     +     +  2  +     +     + --- +  2  +     +     + --- +     +     +     +     +     +  2  +     +     +     +     +
     | V13-B2-CSE7                                                                                |                     |     |     | --- |     |     | --- |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
     | V31-B1-CSE5                                                                                |                     |     | --- |     |     | --- |     |     |     |     |     |     |     |     |     | --- | --- |     |     | --- |     | --- |
     |                                                                                            |                     |     |     |     |     |     |     |     |     | --- |     |     | --- |     |     |     |     |     |     |     | --- |     |
     |                                                                                            |                     |     |     |     | --- |     |     |     |     |     | --- |     |     |     | --- |     |     |     | --- |     |     |     |
   -2+                                                                                            +                     + (2) + (1) + (1) + (1) + (2) + (1) + (2) + (1) + (1) + (1) + (1) + (2) + (2) + (2) + (1) + (1) + (1) + (1) + (1) + (1) + (1) +
-----+--------------------------------------------------------------------------------------------+---------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Measr|+Item                                                                                       |+Rater               | S.1 | S.2 | S.3 | S.4 | S.5 | S.6 | S.7 | S.8 | S.9 |S.10 |S.11 |S.13 |S.14 |S.15 |S.16 |S.17 |S.18 |S.19 |S.20 |S.21 |S.22 |
-----+--------------------------------------------------------------------------------------------+---------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

Fig. 1 All-facet ruler map

The all-facet ruler map is useful for eyeballing initial problems with facets on the logit scale. The first column (labeled "Measurement") is the logit scale, an arbitrary measurement scale centered on 0, running from -2 up to +3 logits. The second column (labeled "Item") contains 39 descriptors, each marked with an ID number, its CEFR level, and its intended CSE level (e.g., V35-C1-CSE8). The descriptors were calibrated in rank order on the logit scale. The third column (labeled "Rater") - with each participant having been assigned an ID number (e.g., T1 for Teacher 1) - shows each participant's ability in discriminating between descriptor levels. The use of levels is demonstrated in the last 21 columns, where S.1 stands for the use of the scales by T1, S.2 for T2, etc.

The second column ("Item") starts between the -2 and -1 logits, with two B1 descriptors (V31 & V11) and one B2 descriptor (V13). Between -1 and 0 logits cluster B1, B1+, and B2 descriptors, with B1+ descriptors overlapping with B1 and B2 descriptors. All the C1 descriptors are higher above 0 logits, with a mixture of one B1 descriptor (V5) and two B2 descriptors (V16 & 34). V5 is judged to be fairly difficult although it is taken from the B1 level. It is evident that participating teachers considered these aspects of vocabulary knowledge to be more demanding than that stated in the CEFR. The spread of the descriptors is around the center of the scale. Descriptors closer to the bottom are considered easier, with those closer to the top more difficult. The logit spread (-1 to +2 logits) indicates that the descriptor difficulty range is narrow, with considerable overlap between B1, B1+, and B2 levels between -1 and 0 logits.

The third column ("Rater") compares the participating teachers with regard to their level of severity/leniency, covering a 5-logit spread (i.e., between -2 and +3 logits). More severe participants appear higher, while more lenient participants lower. Thus, participants T20, T4, T5 and T6 come out with appropriate severity; T18 as extremely lenient.

The 21 columns to the right side of the map show participants' application of the four levels to the descriptors. The all-facet ruler map does not contain information for rater T12. As shown in Fig. 1, T1 did not apply the full level range in rating. T4 assigned lower levels than T3. T18 rarely used B1 level, because the cut-off line between B1 and B1+ was not clear. T10 used all four levels in rating, rating items at B1+, B2, C1 levels more than at B1 level. The cut-offs between B1, B1+ and B2 are not consistent.

In sum, the C1 level descriptors appear to be better distinguished, but considerable overlap may be noted among B1, B1+, and B2 levels, with participants' judgements varying considerably at these levels.

The participant measurement report in Table 3 shows the measures of participant variation in terms of severity/leniency.

The teachers in the first column (labelled "participant") are ordered from the most severe (T12) to the most lenient (T18). Participants' logit value (labelled "Logit") in the second column is shown in descending order, indicating the degree of severity of participants in rating the scales. Participants do not cluster around the center, but are scattered, showing varying severity from around -1.5 to +2.5 logits. According to Papageorgiou (2009), a participant whose logit value is positive "+" is a stricter rater than one whose logit value is negative "-". Two participants (T12 and T1) were very strict in rating the descriptors, with logit values being +2.37 and +2.11 logits respectively. Given that floor and ceiling effects tend to render calibrations outside the central range of -2.0 logit up to +2.0 logit unreliable (North 2000), T12 was removed due to the extreme scores at the C1 level. Three participants (T18, T17 and T2) were extremely lenient, with logit values being -0.97, -0.78 and -0.75. Most participants were in the range of -1 to +1 logits, covering a range of 3 logits. Participants with good model fit were able to discriminate the descriptors well. The standard error in the third column (labelled "S.E.") shows an estimate of the precision of the logit value. Fit statistics - infit and outfit mean square statistics - in the fourth column show the differences in the calibration between expected and observed values. The Infit MnSq (Infit mean square) is "a transformation of the residuals, the difference between the predicted and the observed, for easy interpretation. Its expected value is 1" (Bond and Fox 2007, p. 310). More than 1 indicates misfit, which signals greater variation than expected; less than 1 shows overfit, which indicates less variation than expected: "a result a bit too

Table 3 Participant measurement report: Fit analysis

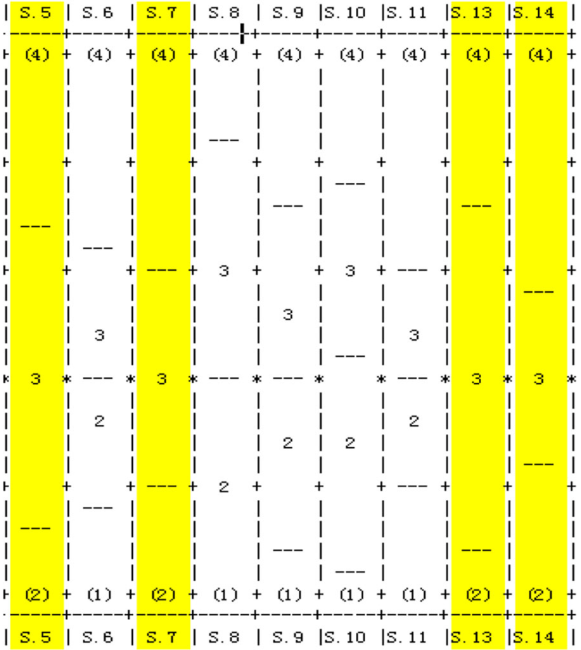
Participant	Logit	S.E.	Infit MnSq
T12	2.37	.54	1.17
T1	2.11	.37	1.33
T7	1.29	.27	1.57
T14	1.13	.25	1.80
T15	1.10	.31	1.47
T21	.52	.21	.96
T13	.34	.29	1.86
T11	.20	.17	.60
T22	.16	.18	.65
T9	.14	.19	.59
T4	.03	.21	.99
T5	.01	.26	2.20
T6	-.01	.17	.44
T20	-.02	.18	1.03
T10	-.15	.21	.63
T8	-.24	.23	.57
T16	-.40	.19	.47
T3	-.44	.19	.63
T19	-.63	.21	.83
T17	-.75	.20	.48
T2	-.78	.20	.48
T18	-.97	.23	.69
Mean	.23	.24	.97
SD	.89	.08	.52

good to be true" (North 2000, p. 231). The mean-square values in the range of 0.5 to 1.5 is conventionally considered indicative of "useful fit" (Weigle 1998). Fourteen participating teachers had infit mean-square fit statistics that were located within the defined fit range. Descriptors with values above 1.5 show misfit. Table 3 shows a considerable amount of misfit among participants, with four participants (T7, T14, T13 and T5) showing heightened degrees of misfit or noisiness (North 2000) - suggesting that these participants' ratings were somewhat unpredictable and inconsistent. Figure 2 displays how participants (especially T5, T7, T13 and T14) used the levels in rating the descriptor scales.

Descriptors with values below 0.5 indicates overfit. The infit mean square values of the following four participants (T6, T16, T17, and T2) were below 0.5 (see Table 3), indicating slight overfit and muted ratings towards a central tendency, although these participants did use the full range of scales in their ratings (see Fig. 3).

FACETS provides a reliability index, with a low reliability index being desired for the participant facet (Linacre 2005). Table 4 denotes differences in leniency among participants, because of the high reliability value of 0.91, and the chi square value ($\chi^2(21) = 183.7$, $p < .00$) was statistically significant at the 1% level. On this basis, the null hypothesis (i.e., there is no variation in participants' discrimination of descriptors) is therefore rejected.

Table 5 summarizes descriptor measurement on the vertical scale. The descriptors and corresponding original levels are arranged in descending logit value order. Infit mean square values indicate the degree to which the scaling of descriptors can be viewed as successful. Wrong scaling can be indicated by examining the logit values in Column 3. For example, V34 is a B2 descriptor but has a logit value of +0.90 which is higher than V2, V4, V39, V18 and V1. These descriptors are C1 descriptors. To elucidate: descriptor V13 was ranked below -1 logit. Descriptor V10 had a similar scaling



```
| S.5 | S.6 | S.7 | S.8 | S.9 |S.10 |S.11 |S.13 |S.14 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
+ (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) + (4) +
|     |     |     |     |     |     |     |     |     |
|     |     |     |     |     |     |     |     |     |
|     |     |     |     |     |     |     |     |     |
|     |     |     | --- |     |     |     |     |     |
+     +     +     +     +     +     +     +     +     +
|     |     |     |     |     | --- |     |     |     |
|     |     |     |     | --- |     |     | --- |     |
| --- |     |     |     |     |     |     |     |     |
|     | --- |     |     |     |     |     |     |     |
+     +     + --- +  3  +     +  3  + --- +     +     +
|     |     |     |     |     |     |     |     | --- |
|     |     |     |     |  3  |     |     |     |     |
|     |  3  |     |     |     |     |  3  |     |     |
|     |     |     |     |     | --- |     |     |     |
*  3  * --- *  3  * --- * --- *     * --- *  3  *  3  *
|     |     |     |     |     |     |     |     |     |
|     |  2  |     |     |     |     |  2  |     |     |
|     |     |     |     |  2  |  2  |     |     |     |
|     |     |     |     |     |     |     |     | --- |
+     +     + --- +  2  +     +     + --- +     +     +
|     | --- |     |     |     |     |     |     |     |
| --- |     |     |     |     |     |     |     |     |
|     |     |     |     | --- |     |     | --- |     |
|     |     |     |     |     | --- |     |     |     |
+ (2) + (1) + (2) + (1) + (1) + (1) + (1) + (2) + (2) +
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| S.5 | S.6 | S.7 | S.8 | S.9 |S.10 |S.11 |S.13 |S.14 |
```

Fig. 2 Misfitting participants

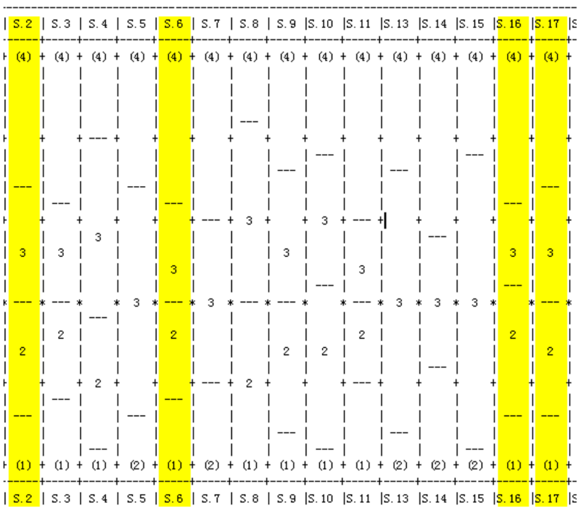


Fig. 3 Overfitting participants

problem in that it was scaled below 0 logits. V26 was judged to be more difficult than V22 and V25. At 0 logits, there were B1 (V21) and B1+ (V23, V26) descriptors. Descriptors V13, V14, V15, V16 were originally taken from the same descriptor (*Can recognize and use a range of vocabulary, idioms, colloquial expressions and technical jargon*), but these appeared at different positions, with V13 at -1 logit value, V14 and V15 at the -0.5 logit value, and V16 at around 0 logit value. All the C1 descriptors clustered between 0 and +2 logits - clearly higher than the other descriptors. Descriptor V34 was ranked as more difficult. Descriptors V5 and V16 appeared at the C1 level, although these were from B1+ and B2 respectively. These descriptors around the center of the scale were ranked as either easier or more difficult.

Out of 39 descriptors, 17 descriptors (V34, V5, V9, V30, V21, V23, V26, V33, V19, V15, V14, V24, V27, V28, V29, V10, and V13) showed higher or lower logit values than those reported for the original CEFR levels.

The acceptable range of infit mean square value for descriptors is 0.4 to 1.2 (Linacre and Wright 1994). Descriptors V35, V3, V6, V16, V32, V36 and V31 have misfit statistics, possibly indicative of greater variation in ratings than expected and suggesting disagreement between ratings. These descriptors do not therefore appear to be sufficiently robust in their descriptions and should not be used as anchor descriptors in future descriptor scaling. An infit mean square below 0.4 indicates overfit. The following four descriptors (V39, V21, V22 and V25) show overfit, suggesting that while ratings varied a little more than expected, generally, there was consistency between participants in their rating of these descriptors. The general consistency suggests that there may well be interdependency between certain descriptors.

Table 4 Participant measurement report: Differences in participants

Statistics	Results
Reliability	0.91
Chi-square	183.7
d.f.	21
Significance	0.00

Table 5 Descriptor measurement report

Descriptor	Level	Logit	S.E.	Infit MnSq
V35	C1-CSE8	1.90	.39	1.85
V3	C1-CSE8	1.18	.31	1.67
V7	C1-CSE8	1.18	.31	1.16
V20	C1-CSE8	1.18	.31	1.17
V6	C1-CSE8	1.08	.31	2.20
V8	C1-CSE8	1.08	.31	1.04
V34	B2-CSE7	.90	.30	.45
V2	C1-CSE8	.73	.29	.85
V4	C1-CSE8	.73	.29	.73
V39	C1-CSE8	.73	.29	.33
V18	C1-CSE8	.34	.28	.85
V1	C1-CSE8	.26	.27	.83
V5	B1 + -CSE6	.26	.27	.64
V9	C1-CSE8	.26	.27	1.05
V30	C1-CSE8	.19	.27	.86
V16	B2-CSE7	.11	.27	1.66
V21	B1-CSE5	.04	.27	.35
V23	B1 + -CSE6	.04	.27	.43
V26	B1 + -CSE6	-.04	.27	.69
V22	B2-CSE7	-.11	.28	.32
V33	B1 + -CSE6	-.11	.28	.42
V19	B1 + -CSE6	-.26	.28	.65
V25	B2-CSE7	-.26	.28	.35
V17	B2-CSE7	-.34	.28	.76
V15	B2-CSE7	-.42	.28	.45
V38	B1 + -CSE6	-.42	.28	1.19
V14	B2-CSE7	-.50	.28	.48
V24	B1-CSE5	-.58	.29	.61
V27	B1-CSE5	-.58	.29	.45
V12	B1 + -CSE6	-.66	.29	.55
V28	B2-CSE7	-.66	.29	.54
V29	B2-CSE7	-.66	.29	.98
V32	B1-CSE5	-.66	.29	1.22
V37	B1-CSE5	-.66	.29	.85
V36	B1-CSE5	-.75	.29	1.39
V10	B2-CSE7	-.84	.30	.80
V11	B1-CSE5	-1.02	.31	1.03
V13	B2-CSE7	-1.22	.32	.58
V31	B1-CSE5	-1.44	.34	1.42
Mean		.00	.29	.87
SD		.77	.02	.45

Discussion

The results of the data analysis indicated that considerable variation exists in scaling the CSE vocabulary descriptors to those of the CEFR.

The data analysis shows that the top level - C1-CSE8 - was clearly separated, although two descriptors (V34 and V5) were varied more greatly than expected (see Table 6). Descriptors at B1-CSE5, B1 + -CSE6 and B2-CSE7 were ranked in a mixed order of difficulty. A few descriptors at levels of B1-CSE5 (V21, V24, V27) and B1 + -CSE6 (V5, V23, V26, V33, V19) were ranked as higher levels, whereas several descriptors at levels of B2-CSE7 (V15, V14, V28, V29, V10, V13) as lower levels. There was considerable overlap between B1+ and B1, B2 levels, suggesting that participants had difficulty in scaling these descriptors.

Of the 39 descriptors, there were 17 problematic descriptors (V34, V5, V9, V30, V21, V23, V26, V33, V19, V15, V14, V24, V27, V28, V29, V10, and V13), which ranged across more than three levels. These problematic descriptors were all rated one level higher or one level lower than the original CEFR levels. Descriptors V33 and V34, denoting size of vocabulary knowledge, were along with V19, V24, and V27, ranked one level higher

Table 6 Descriptors with scaling problems

Logit	ID	CEFR-CSE level	Rated level	Descriptor wording
.90	V34	B2-CSE7	C1-CSE8	Knows about 6,000 words, and words from the AWL (Academic Word List) and 1,200 phrases, among which 2,200 are active words
.26	V5	B1 + -CSE6	C1-CSE8	Knows many most frequently used idioms
.26	V9	C1-CSE8	B2-CSE7	Knows expressions based on polysemy
.19	V30	C1-CSE8	B2-CSE7	Finds several vocabulary options in almost all speaking and writing situations
.04	V21	B1-CSE5	B2-CSE7	Can produce synonyms of basic words in different contexts
.04	V23	B1 + -CSE6	B2-CSE7	Can produce the synonyms to the most common words of most parts of speech
-.04	V26	B1 + -CSE6	B2-CSE7	Can produce the antonyms to the most common words of most parts of speech
-.11	V33	B1 + -CSE6	B2-CSE7	Knows the meaning of about 5,000 words and words from the AWL (Academic Word List), and 700 phrases, among which 2,000 are active words
-.26	V19	B1 + -CSE6	B2-CSE7	Can use a range of affixations to produce basic words
-.42	V15	B2-CSE7	B1 + -CSE6	Can recognize and use a range of colloquial expressions
-.50	V14	B2-CSE7	B1 + -CSE6	Can recognize and use a range of idioms
-.58	V24	B1-CSE5	B1 + -CSE6	Can produce antonyms of basic words in different contexts
-.58	V27	B1-CSE5	B1 + -CSE6	Can produce some frequent collocations
-.66	V28	B2-CSE7	B1 + -CSE6	Can express meanings by adding affixation to familiar words, e.g., have a *review*
-.66	V29	B2-CSE7	B1-CSE5	Has a good command of vocabulary related to everyday situations
-.84	V10	B2-CSE7	B1-CSE5	Knows a number of principles of word formation, e.g., agree - agree*able*
−1.22	V13	B2-CSE7	B1-CSE5	Can recognize and use a range of vocabulary

than their original CEFR levels. Descriptors V5 and V21 were ranked two levels higher than those in the original CEFR levels. Descriptors V9 and V30 were scored one level lower at B2 level, and V15, V14, V28 were also scored one level lower at B1+ level. Descriptors V29, V10, V13 - at B2 level - were rated two levels lower than their original CEFR levels. V26 and V23 - interrelated descriptors - were originally from the same CEFR level (B1 + -CSE6), but were ranked one level above.

Participants' judgements resulted in problematic misfitting and overfitting descriptors, among which there were qualitative and quantitative descriptors (see Table 6). Possible reasons for mismatch might be due to participants having difficulty in distinguishing such terms as “a wide range of”, “a large number of” and “a range of”. Take descriptor V34, for example. This descriptor (*knows about 6,000 words, and words from the AWL (Academic Word List) and 1,200 phrases, among which 2,200 are active words*) had a comparatively high logit value (0.90), and was rated as more difficult than its original B2 level. According to the CECR, “Has a command of a lexical repertoire of 7,675 words and 1,870 phrases” is the requirement for advanced users. So the participants rated the descriptor one level higher. It is also likely that the vocabulary size stipulated in the CECR has greater discrepancy in comparison with the CEFR. Descriptor V5 was also rated as more difficult than its CEFR level. Descriptor V21 showed overfit, with an infit mean square of 0.35. The most problematic scaling lay with the B1+ level in that there were different degrees of overlap between B1+ and B1, B1+ and B2 level.

It is likely that the closeness of the intervals between B1, B1+, and B2 led to the overlap between these levels.

Table 7 summarizes discrepancies between participants' ratings and the different levels of descriptor.

Among the 39 descriptors, seven misfitting descriptors showed that there were considerable differences between participants' ratings and the original CEFR levels, indicating a big variation in ratings. Level disagreement was one reason for misfit. For example, two teachers rated V35 as B1+, two teachers rated it as B2, and the remaining 17 teachers rated it as C1. V36, with a logit value of -0.75, belonged to B1 level, but was ranked higher above its original CEFR level.

There were four descriptors whose logit values showed overfit, as can be seen from Table 8.

The overfit in Table 8 might be attributed to two issues. The first is a lack of variation in judgement-making, indicating more consistency between participants in rating descriptors; the second is descriptor interdependency (McNamara 1996). The overfit for V22 and V25, for example, can be attributed to interdependency. These two descriptors were originally from the same B2 descriptor (*Can produce synonyms and antonyms of most common words in different contexts*), but were ranked by participants as descriptors at different levels. The logit value of V22 was -0.11, while the logit value of V25 was -0.26. The reason for this might be because participants perceived "synonyms" as more difficult than "antonyms". As for V39 and V21, the variation in judgement-making was a little greater than expected.

Conclusion

This paper has reported a pilot study with the aims of constructing CSE descriptor scales for College English vocabulary education in China. The study has provided an objective measurement from teachers' subjective judgements of the CEFR-based CSE vocabulary descriptor scales. A Multi-Faceted Rasch Measurement analysis of the data indicated that considerable variation existed among teachers in the attempt to align the CSE descriptors with those of the CEFR. The results present a mixed picture, with some scales matching, and others not.

While participants were familiar with CECR descriptors, they were less familiar with those of the CEFR, in particular descriptors at the B levels. On the basis of the 39

Table 7 Reasons for item misfit

ID	Descriptor wording	Level	Reason for misfit
V35	Knows the meaning of about 8,000 words and words from the AWL (Academic Word List), and 2,000 phrases	C1-CSE8	Level disagreement
V3	Understands and uses a wide range of technical jargon	C1-CSE8	Level disagreement
V6	Knows quite a large number of less usual idioms	C1-CSE8	Level disagreement
V16	Can recognize and use a range of technical jargon	B2-CSE7	Level disagreement
V32	Knows the meaning of 400–500 idioms or fixed collocations	B1-CSE6	Level disagreement
V36	Can recognize the meaning of 1,500–2,000 most frequent everyday vocabulary related to a range of basic personal and familiar situations	B1-CSE6	Level disagreement
V31	Knows the meaning of 3,000 words	B1-CSE6	Level disagreement

Table 8 Reasons for item overfit

ID	Descriptor wording	Level	Reason for overfit
V39	Has a good command of over 5,000 words	C1-CSE8	Level agreement
V21	Can produce synonyms of basic words in different contexts	B1-CSE5	Level agreement
V22	Can produce synonyms of most common words in different contexts	B2-CSE7	Interdependency
V25	Can produce antonyms of most common words in different contexts	B2-CSE7	Interdependency

descriptor scaling, more descriptors should be collected and categorized to enrich the current pool of descriptors. Further, more representative tasks should be designed in relation to each CSE level to enable participants to have a better understanding of the descriptive scales. In addition to the quantitative study, qualitative follow-up interviews should also be conducted to investigate in greater depth participants' perceptions with relation to their judgement-making.

The current study constitutes external validation of participating teachers' judgements of how well the CECR vocabulary descriptors align with those of the CECR. In light of the MFRM analysis, it can be concluded that descriptors at the top level (i.e., C1 in the CEFR/CSE 8 in the CECR) were clearly separated and ranked as expected while the most problematic overlapping scaling lay with the intermediate levels (i.e., B1, B1+, and B2/CSE 5, CSE6, and CSE 7) - indicating the difficulty that participants experienced in ranking the intermediate levels.

As with many projects, the current study has its limitations. Pressure of time and resources limited the number of participants and their familiarization with the descriptors. Although participating teachers were given a brief introduction to the CEFR, it is quite likely that they were not adequately familiar with the CEFR levels, in particular B1, B1+, and B2 levels. More training might be needed to familiarize participants with the CEFR in order to have sufficient understanding and detailed knowledge of the CEFR vocabulary descriptors so that they can reach a better agreement on the scales. Further, the findings of the study suggest that participant judgment alone was not sufficient. Multiple sources of evidence should also be provided to triangulate the empirical evidence so that consistent interpretation and modification can be provided.

While the study offered a manageable approach for calibrating the CEFR against the CSE for tertiary English education in China, further in-depth teacher training with a set of more experienced teachers is recommended for conducting a future external validation study since it is somewhat unlikely that a single one-off study will provide sufficient evidence of alignment (Martyniuk 2010). Iterative cycles of testing and revision should be provided to develop more comprehensive illustrative vocabulary descriptors representative enough to adequately reflect the range of CSE illustrative scales.

As Kaftandjieva (2004) notes, the CEFR scales are valid, but this does not guarantee that the scales will be validly interpreted as standards in all contexts in which they may be used. The current study has shown that validation evidence of the CSE vocabulary scales can be provided as a reference point when scale values of the descriptors and participant agreement are calculated. Participants' informed judgements are important to establish the validity of the CSE scales in aligning them with the CEFR, with the CSE vocabulary descriptors continuing to be expanded to enrich the CSE descriptive and illustrative scales.

Appendix 1

Table 9 The Vocabulary Descriptor Bank

CEFR-CSE level	ID	Descriptor wording
C1-CSE8	V1	Understands and uses a wide range of vocabulary
C1-CSE8	V2	Understands and uses a wide range of idioms
C1-CSE8	V3	Understands and uses a wide range of technical jargon
C1-CSE8	V4	Understands and uses a wide range of colloquial expressions
B1 + -CSE6	V5	Knows many most frequently used idioms
C1-CSE8	V6	Knows quite a large number of less usual idioms
C1-CSE8	V7	Knows the synonyms of less common words
C1-CSE8	V8	Knows the antonyms of less common words
C1-CSE8	V9	Knows expressions based on polysemy
B2-CSE7	V10	Knows a number of principles of word formation, e.g., agree - agree*able*
B1-CSE5	V11	Knows how to use basic word formation principles
B1 + -CSE6	V12	Can recognize polysemy, e.g., *back* a car/*back* a proposal
B2-CSE7	V13	Can recognize and use a range of vocabulary
B2-CSE7	V14	Can recognize and use a range of idioms
B2-CSE7	V15	Can recognize and use a range of colloquial expressions
B2-CSE7	V16	Can recognize and use a range of technical jargon
B2-CSE7	V17	Can use a number of frequently used idioms
C1-CSE8	V18	Can use words idiomatically and appropriately
B1 + -CSE6	V19	Can use a range of affixations to produce basic words
C1-CSE8	V20	Can use affixations even in the case of unusual and abstract words to form even less common expressions
B1-CSE5	V21	Can produce synonyms of basic words in different contexts
B2-CSE7	V22	Can produce synonyms of most common words in different contexts
B1 + -CSE6	V23	Can produce the synonyms to the most common words of most parts of speech
B1-CSE5	V24	Can produce antonyms of basic words in different contexts
B2-CSE7	V25	Can produce antonyms of most common words in different contexts
B1 + -CSE6	V26	Can produce the antonyms to the most common words of most parts of speech
B1-CSE5	V27	Can produce some frequent collocations
B2-CSE7	V28	Can express meanings by adding affixation to familiar words, e.g., have a *review*
B2-CSE7	V29	Has a good command of vocabulary related to everyday situations
C1-CSE8	V30	Finds several vocabulary options in almost all speaking and writing situations
B1-CSE5	V31	Knows the meaning of 3,000 words
B1-CSE5	V32	Knows the meaning of 400–500 idioms or fixed collocations
B1 + -CSE6	V33	Knows the meaning of about 5,000 words and words from the AWL (Academic Word List), and 700 phrases, among which 2,000 are active words
B2-CSE7	V34	Knows about 6,000 words, and words from the AWL (Academic Word List) and 1,200 phrases, among which 2,200 are active words
C1-CSE8	V35	Knows the meaning of about 8,000 words and words from the AWL (Academic Word List), and 2,000 phrases
B1-CSE5	V36	Can recognize the meaning of 1,500–2,000 most frequent everyday vocabulary related to a range of basic personal and familiar situations
B1-CSE5	V37	Can produce the meaning of 1,500–2,000 most frequent everyday vocabulary related to a range of basic personal and familiar situations
B1 + -CSE6	V38	Can produce the meaning of 2,000 most frequent words
C1-CSE8	V39	Has a good command of over 5,000 words

Acknowledgements

We would like to thank Professor JIN Yan and Professor YU Guoxing for their valuable suggestions on earlier drafts of the manuscript.

Funding

This paper was based on the project "The Applicability of the CEFR for English Language Education in China" (GZ20140100) funded by Foreign Language Teaching and Research Press and the Key Project of the National Social Science "Corpus-based Criterial Featured Parameter Study of College English Language Competence" (13AYY005) funded by the National Planning Office of Philosophy and Social Science.

Authors' contributions

WZ, as a group leader, participated in the Common Chinese Framework of Reference for English (CCFR-E) project funded by Foreign Language Teaching and Research Press. WZ drafted the manuscript; the other authors, in particular DC, provided insight comments on the draft. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Northeastern University, Shenyang, China. [2]The Education University of Hong Kong, Hong Kong, China.

Received: 4 November 2016 Accepted: 14 February 2017
Published online: 28 February 2017

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Macmillan.

Alderson, J. C. (Ed.). (2002). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Case studies*. Strasbourg: Council of Europe.

Alderson, J. C. (Ed.). (2004). *The shape of things to come: Will it be the normal distribution?* Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Alderson, J. C., Kuijper, H., Nold, G., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly, 3*(1), 3–30.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the Human Sciences* (2nd ed.). London: Lawrence Erlbaum Associates.

Finch, A. (2009). Europass and the CEFR: Implications for language teaching in Korea. *English Language and Literature Teaching, 15*(2), 71–92.

Han, B. (2006). A review of foreign language proficiency scales. *Foreign Language Teaching and Research, 38*(6), 443–450.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689–696.

Huhta, A., & Figueras, N. (2004). Using the CER to promote language learning through diagnostic testing. In K. Morrow (Ed.), *Insights from the Common European Framework* (pp. 65–76). Oxford: Oxford University Press.

Kaftandjieva, F. (2004). *Standard setting. Section B 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*. Strasbourg: Council of Europe.

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* (pp. 106–129). Strasbourg: Council of Europe.

Linacre, J. M. (2005). *FACETS Rasch measurement computer program version 3.58.*. Chicago: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Martyniuk, W. (Ed.) (2010). Aligning tests with the CEFR: Refelections on using the Council of Europe's draft Manual (Vol. 33). Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. Harlow: Longman.

Ministry of Education. (2006). *English Curriculum for Senior Secondary Education*. Beijing: Beijing Normal University.

Ministry of Education. (2007). *College English Curriculum Requirements*. Beijing: Higher Education Press.

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks* (Vol. 36, pp. 135–163). Cambridge: Cambridge University Press.

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

North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal, 91*(4), 656–659.

North, B. (2014). *The CEFR in Practice*. Cambridge: Cambridge University Press.

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

Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Oxford: Peter Lang.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*, 139–152.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.

Wu, J. (2012). Policy perspectives from Taiwan. In M. Byram (Ed.), *The Common European Framework of Reference: The globalisation of language education policy (Vol. 23): Multilingual matters*.

Yang, H., & Gui, S. (2007). On developing a unified Asian framework of English rating scales. *Foreign Languages in China, 4*(2), 34–37. 67.

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