English Language Teaching in China: Developing Language Proficiency Frameworks

24

Jianda Liu and Mingwei Pan

Contents

Abstract

This chapter introduces the development of an English language proficiency scale for English language teaching in the Chinese context. Based on the Communicative Language Ability model (Bachman (1990) Fundamental considerations in language testing. Oxford University Press, Oxford; Bachman and Palmer (1996) Language testing in practice: designing and developing useful language tests. Oxford University Press, Oxford) and the use-oriented principle, this scale

J. Liu (✉)
Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China
e-mail: jackliu@gdufs.edu.cn

M. Pan
Shanghai International Studies University, Shanghai, China
e-mail: mwpan@shisu.edu.cn

X. Gao (ed.), *Second Handbook of English Language Teaching*, Springer International Handbooks of Education, https://doi.org/10.1007/978-3-030-02899-2_28

defines English language competence for English language learners in the Chinese context, with specific reference to the teaching of English. According to the scale, English language competence consists of comprehension competence and expression competence. Unique to the Chinese context, the scale developers also incorporated into the framework the ability to mediate between English and Chinese, i.e., translation competence and interpretation competence. The English proficiency scale includes sets of "can-do" statements that define different standards for English learners at different levels. This chapter also highlights the rationale and characteristics of the proficiency scale and its significant implications for English language learning and teaching in China.

Keywords
Language proficiency scales · Comprehension competence · Expression competence · Mediation ability

Introduction

Language proficiency scales are intended to describe the extent to which language learners or users at different proficiency levels can use the target language in real-life situations (North 2000). They may serve different purposes and play an important role in integrating language assessment with learning and teaching (Alderson 2005).

Back in the 1950s, when English teaching and learning played a primary role in the military and government arenas, the Foreign Service Institute Scale (FSI) and the Interagency Language Roundtable Scale (ILR) were developed with a view to recruiting and appraising soldiers and civil servants and coordinating various US federal agencies to keep abreast of modern methods and technology (see more details at http://www.govtilr.org/skills/ILRscale2.htm). Since that time, however, the focus of language proficiency scales has shifted more toward the area of language education. In the 1980s the American Council on the Teaching of Foreign Languages (ACTFL; see more details at https://www.actfl.org) stipulated fine-grained descriptors with regard to the basic skills of English learning. Since the inception of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001; North 2000), language proficiency scales have increasingly far-reaching effects on global language policymaking, curriculum design, teaching material development, and language (Figueras 2012). As a result, the development of nation- or region-wide English proficiency scales continued apace, such as the Canadian Language Benchmarks (CLB; Center for Canadian Language Benchmarks 2012, 2015) in Canada, the International Second Language Proficiency Ratings (ISLPR; see Wylie and Ingram (2010) for more details) in Australia, and the CEFR-J in Japan (see more details at http://www.tufs.ac.jp/ts/personal/tonolab/cefr-j/english/whatis.html).

China, which boasts the largest population of English learners worldwide, has not been immune to the sweeping impact of language proficiency scales. There has been an increasing awareness of scaling English learners' competence, and researchers,

practitioners, and policymakers in China all agree that a unified proficiency scale is urgently needed to describe learners' performance and streamline their competence across different educational stages and different regions in the Chinese context (Dai and Zhang 2001; Yang and Gui 2007). In 2014 the State Council of China issued a document entitled "Deepening the reforms on educational exams and recruitment systems." One pressing task, as highlighted in the document, was to develop an English language proficiency scale for English learners and users across different proficiency levels in China. In this context the National Education Examinations Authority (NEEA), endorsed by the Ministry of Education, China, initiated a nationwide project to develop an English language proficiency scale, known as China's Standards of English Language Ability (CSE). This set out to (1) define and describe the English competencies that learners at different educational phases are supposed to reach; (2) provide references and guidelines for English learning, teaching, and assessment; and (3) enrich the existing body of language proficiency scales for future alignment on a global basis (Liu 2015a).

With a review of the extant literature on language proficiency scales as a point of departure, this chapter describes how the CSE was developed based on the Communicative Language Ability (CLA) model (Bachman 1990; Bachman and Palmer 1996, 2010) and the use-oriented principle. Light is also shed on how descriptors were categorized and scaled and how the scales were validated. Finally, guidance is provided on how exemplar activities, as by-products of the CSE, may inform language teaching and assessment.

Theories and Principles of CSE Development

In order to justify the rationale of the CSE, we will review the well-established prevailing language proficiency scales outside China, the models or frameworks concerning language ability, and the use-orientedness of the CSE. We will also present an operationalizable framework for the development of the CSE.

Language Proficiency Scales Outside China

At the beginning of the development of the CSE, a number of existing language proficiency scales, such as the CEFR and the CLB, were reviewed. Among these scales it was found that the CEFR stood out as the most influential in the development and validation of the CSE, for the following reasons. First, the CEFR can be regarded as one of the most influential existing language proficiency scales. Many proficiency scales are adaptations or (sub)-branches of the CEFR, for example, the CEFR-J. Many international testing batteries or organizations also align their tests or proficiency scales with the CEFR, such as the ILR, the ACTFL, and the CLB, as shown in Table 1. Second, the CEFR is innovative in the sense that it incorporates the collaborative co-construction of meaning, as well as plurilingual and pluricultural competence (North and Panthier 2016). This innovation of the CEFR has

Table 1 Alignment of other scales to the CEFR

CEFR	ILR	ACTFL	CLB
A1	0/0+/1	Novice (low/mid/high)	1/2
A2	1+	Intermediate (low/mid/high)	3/4
B1	2	Advanced low	5/6
B2	2+	Advanced mid	7/8
C1	3/3+	Advanced high	9/10
C2	4/4+/5	Superior	11/12

attracted much interest in multicultural language education contexts. Third, the CEFR is regularly updated; the latest version was released around the end of 2017 (Council of Europe 2018).

In general, the CEFR has been developed to provide a common basis for language syllabi, curriculum guidelines, examinations, and textbooks and to relate a European credit scheme to fixed points in a framework (van Ek 1975). The global scaling of this framework was largely inspired by commonly referenced proficiency levels in other documents, such as *Threshold*, *Vantage*, *Waystage*, *Breakthrough*, *Effective Operational Proficiency*, and *Mastery* (Alderson 2002), which correspond to the proficiency levels of A1, A2, B1, B2, C1, and C2, respectively. It was then developed with detailed descriptors for each level, which includes benchmarked behavioral characteristics in various domains (Little 2006). The Council of Europe (2001) claims that the CEFR is comprehensive as "it should attempt to specify as full a range of language knowledge, skills and use as possible . . . and all users should be able to describe their objectives, etc. by reference to it" (p. 7).

Nevertheless, the CEFR is not without its critics. First, the construct of language ability as reflected in the CEFR or its descriptors is basically drawn from teachers' perceptions only, while learners' or other stakeholders' perceptions were not included. An overreliance on quantitative methods alone in dealing with teacher's perceptions might also be problematic. In addition, the descriptors take "insufficient account of how variations in terms of contextual parameters may affect performances by raising or lowering the actual difficulty level of carrying out the target 'can-do' statement" (Weir 2005a, p. 281). Second, the CEFR descriptors seem to lack specificity. At certain points the descriptors may seem redundant and reader-unfriendly (Alderson 2010). For example, a B1 descriptor, "Can give a prepared straightforward presentation on a familiar topic within his/her field which is clear enough to be followed without difficulty most of the time, and in which the main points are explained with reasonable precision," is embedded with various constraints from different perspectives, such as quality of presentation and addresser/addressee of presentation, so that users may be confused about the real foci of the descriptor per se. Third, while the CEFR claims to cover aspects of both proficiency and development in its six ascending levels of proficiency, it fails to do so consistently (Alderson et al. 2006; Hulstijn 2011; Norris 2005). Thus, a host of researchers (e.g., Cumming 2009; Fulcher 2004; Hulstijn 2007; Spolsky 2008) have expressed concerns regarding the rationale of the CEFR. For instance, Spolsky (2008) criticizes

the CEFR as "arbitrary" standards designed to produce uniformity, while Cumming (2009) points out the dilemma of the imprecision of standards such as the CEFR "in view of the complexity of languages and human behaviour" (p. 92). Fourth, although the CEFR developers tried to include translation and interpretation abilities into the new version, whether they have been able to accomplish this seems uncertain, given the resistance and criticism from the field of translation studies.

Models of Language Ability

As reviewed above, it appears the CEFR was intended to streamline language proficiency levels across different social and educational contexts in Europe. However, in China, where English mainly plays the role of a foreign language, CSE developers cannot directly adapt the CEFR (see Zou et al. 2015 for more details). A new proficiency scale applicable to the Chinese context should be prioritized. Before its development, however, CSE developers also considered the language ability models that were indicative of the theoretical underpinnings of the CSE.

Research into the construct of language ability dates back to the 1960s, when Lado (1961) and Carroll (1961, 1968) published their interpretations of language ability. The structuralist approach to language and language ability (e.g., Lado 1961) was followed by a discussion about linguistic competence versus performance and the definition of "communicative competence" (Hymes 1972, 1973, 1982; Halliday 1973, 1976). In the 1980s Canale and Swain (1980; Canale 1983) proposed the first componential model of communicative competence, which was further extended into various pedagogical adaptations of communicative competence (e.g., Celce-Murcia et al. 1995; Savignon 1983). Similarly, Douglas (2000) proposed a model with a particular view of language use for specific purposes, where professional or topical knowledge is equally emphasized (see Purpura 2008 for a detailed review of how language ability models have evolved).

In an eclectic collection of and critique on the above conceptualizations, Bachman (1990) and Bachman and Palmer (1996) posited the Communicative Language Ability (CLA) model, where language ability was perceivably constructed as "consisting of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use" (Bachman 1990, p. 84). The CLA model not only inherits organizational competence in its traditional sense but also embeds strategic competence and regards it as not just serving a compensatory function, which, to a certain extent, alludes to Canale's (1983) refined model. More importantly, it also recognizes the roles of cognitive strategies and pragmatic competence, together with their impact on the realization of communicative competence.

On the whole, this model is theoretically sound and has been empirically validated to a certain extent (though pragmatic competence has not been vigorously validated) and is merited as a state-of-the-art representation (Alderson and Banerjee 2002). Consequently, CSE developers adopted a modified version of the CLA model by incorporating ideas of cognitive abilities, which largely derive from Weir's

(2005b) socio-cognitive framework. As such, the CSE developers referred to the CLA model to a certain degree, but they also operationalized it for scale development purposes, to be elaborated below.

Use-Orientedness

Contingent upon different purposes, language proficiency scales correspond to different orientations. The CEFR claimed to be action-oriented, where more emphasis is laid on whether and, if so, how language users integrate various language skills in performing particular activities in social and public settings – i.e., performing different functions with different texts and topics in different language activities. The CEFR adopts illustrative descriptor scales to highlight users' performance in receptive, productive, and interactive language activities. This orientation has been proven to be appropriate, given the role that English plays in the European context. Europeans, most of whom use English as a second language, are supposed to communicate in the language, either spoken or written. Although English learners and users in China are more likely to learn English in an educational context, they are encouraged to use it in different domains. Thus, in order to justify the real use of English as a target language (Bachman and Palmer 2010), prioritizing how English is used in the Chinese context is a guiding principle of CSE development (Yang 2015).

Framing an Operationalizable Proficiency Scale

In the context of the above review, CSE developers formulated an operationalizable framework. As illustrated in Fig. 1, the core of the CSE reflects an overarching notion of language ability, within which language competences and strategies cofunction in performing a language activity. This mechanism sits comfortably within the CLA model, where communication success depends upon the language competences learners and users resort to, as well as the strategies they employ in an activity.

In a componential sense, language ability can be further divided into comprehension (listening and reading) and expression (speaking and writing). In language comprehension, learners/users are supposed to remember/understand, apply/analyze, or evaluate/create texts (Anderson and Krathwohl 2001), with an ascending order of cognitive demand. In a similar vein, learners/users are expected to convey, explain, or persuade. In actual performance, both comprehension and production channels are enabled and interact with required pragmatic knowledge for communication effectiveness. In addition, mediation is also considered as part of language competence, referring to the fact that language learners/users resort to comprehension and expression in mediating activities.

In this scale, only translation and interpretation are considered. Although it is not a direct component (the dotted-line square) as it involves translation and interpreting

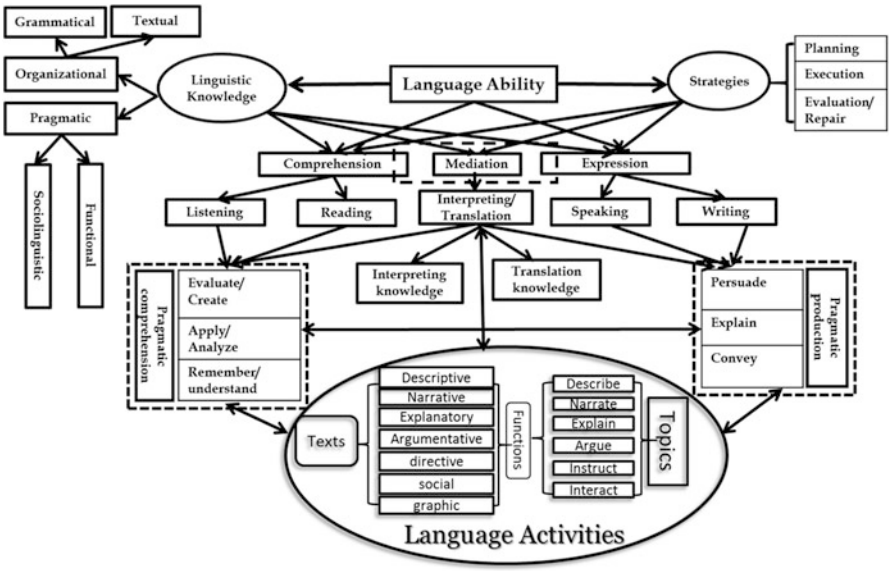

Fig. 1 An operationalizable framework for the CSE

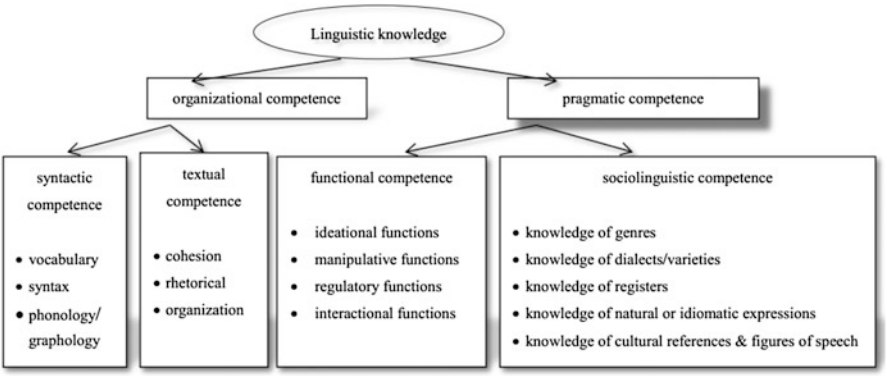

Fig. 2 Linguistic knowledge in the CSE

knowledge that is not linguistically laden, this sub-ability functions in a similar fashion when compared with comprehension and expression.

In order to realize their language competence, learners/users also need to employ their linguistic knowledge, which consists of organizational competence and pragmatic competence. The former can be further broken down into syntactic competence and textual competence; the latter includes functional competence and sociolinguistic competence (see Fig. 2; Bachman 1990).

Apart from language competence, strategies are also involved, which can be divided into planning, execution, and appraising/compensation (Council of Europe 2001).

Different language sub-abilities are heavily involved in all of them. It is noteworthy that the strategies related to different language sub-abilities vary in terms of their names. For example, within the umbrella term of appraising/compensation, the specific name for the writing sub-scales might be editing/proofreading, while that for the speaking sub-scales would be repairing.

As shown in Fig. 1, in congruence with the different functions that communication mainly serves, different sub-abilities deal with a plethora of texts, including narrative, descriptive, expositive, argumentative, directive, social, and graphic texts (Jackson and Stockwell 2011) to narrate, describe, expose, argue, instruct, and interact on different topics. Be it comprehension, expression, or mediation, learners'/users' language competence is ultimately realized in language activities in relation to the abovementioned text types and functions.

Therefore, in order to streamline the framework across different sub-ability scales, the CSE developers laid down a "three-fold four-layer hierarchical framework." Being "threefold" means that language ability is described from three perspectives: language competences, linguistic knowledge, and strategies. Being "four-layer" means that when language competences are described, the descriptors are structured in a hierarchical system. Language ability is the top layer, beneath which there are language comprehension, language expression, and mediation. The third layer is subdivided with global competence descriptors for different sub-abilities. When sub-abilities are instantiated by an assortment of texts with different functions, descriptors specific to sub-abilities constitute the fourth layer (Liu and Han 2018).

Developing and Validating the CSE

Having reviewed the existing competence scales, models of language ability as well as an operationalizable framework for CSE development, in this section we will detail the nuts and bolts of the development and validation of the CSE.

The Development Procedure

Figure 3 outlines the CSE development procedure. As illustrated in the figure, the CSE development can be divided into three major phases. The first phase primarily deals with collecting descriptors, which derived from not only a wealth of literature but also from exemplar generation (described in more detail below). In the second phase, based on expert and teacher judgments, the CSE developers conducted trial validation on a working group basis. During this process the developers removed duplicate descriptors, blended similar descriptors, and categorized descriptors into an operationalizable framework for the CSE, as discussed above. The last phase was composed of two field studies to finalize the scaling. In the first field study, all the descriptors were randomly spread into different sets of questionnaires, which were administered to language education experts and frontline teachers as well as learners/

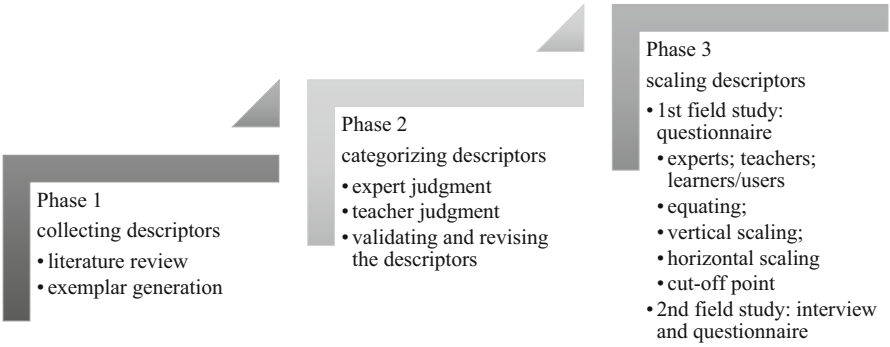

Fig. 3 The CSE development procedure

users. The participants were asked to report the extent to which their students (if the participants were teachers) or they themselves (if they were learners/users) could perform in relation to each descriptor. Based on the results, statistical analyses were conducted to determine the cutoff points for each proficiency level. The second field study, which was smaller in scale, tried to elicit the responses of teachers of various educational stages to the same descriptors, so that horizontal scaling could be carried out for the calibration of the cutoff points.

For convenience of initial scaling, the CSE descriptors were first categorized into nine working levels, CSE 1 to CSE 9, which were arranged in ascending order from lower proficiency levels to higher ones. The descriptors at lower proficiency levels were assumed to be within the competence range of higher proficiency levels. Similarly, to facilitate the trial validation, each level was tentatively aligned to a particular group of English learners in China: CSE 1 corresponded roughly to Grade 3 primary school pupils (a starting point at which most, if not all, English learners have received formal EFL instruction for 1 year), CSE 2 to primary school leavers (or Grade 6 primary school pupils), CSE 3 to junior high school graduates, CSE 4 to senior high school graduates, CSE 5 to non-English major sophomores, CSE 6 to non-English major undergraduates or English major sophomores, CSE 7 to English major undergraduates, CSE 8 to English major postgraduates, and CSE 9 to professional users such as professional translators or interpreters.

The CSE development was a huge project that spanned a long period of time, with a total of 8 working groups including about 200 language assessment experts and PhD students involved. The eight groups were responsible for listening, reading, speaking, writing, translation, interpretation, organizational knowledge, and pragmatic knowledge, respectively. To facilitate inter- and intragroup communication, an online platform was also specifically designed for this project. It should be noted that, except for organizational knowledge and pragmatic knowledge, all the other groups' sub-scales were inclusive of both competences and strategies. In order to facilitate intergroup communication, two special working groups, namely, the coordination group and the validation group, were in charge of logistical issues and data processing, respectively.

Developing the CSE

In order to provide a clearer picture of the research methods, this part provides more detail about how the CSE descriptors were collected and revised during the first two phases of the research design outlined above.

The initial CSE descriptor pool was huge, containing descriptors from various sources ranging from the most common language proficiency scales at home and abroad to curricula, teaching syllabi, textbooks, workbooks, and test syllabi and specifications (Liu 2015b). Apart from searching the literature, the developers also used exemplar generation to supplement the collection method. In this, the participants, mainly frontline teachers external to the project, were asked to produce new language competence descriptors by referring to sample ones. This proved to be effective because there was a scarcity of top- and bottom-level descriptors, and the newly generated descriptors by primary school teachers and postgraduate supervisors enriched the descriptor pool. In addition, the translation and interpretation working groups did not have much literature to refer to concerning their specific sub-abilities. Contributions from translation and interpretation practitioners again significantly expanded the descriptor pool. In total, the first CSE descriptor pool contained 16,477 descriptors.

However, a large descriptor pool does not necessarily mean a pool of high-quality descriptors. After collecting the descriptors, the CSE developers conducted two rounds of screening and revision, where all the descriptors were cross-checked on an intra- and intergroup basis. Apart from this large-scale cross-check, each working group also conducted rounds of workshops to elicit teachers', students', and parents' responses to the raw descriptors so that more feedback could be collected and used for reference in revising the descriptors.

Three guidelines consistently threaded through the whole process of descriptor screening and revision. First, each descriptor must take the form of a "can-do" statement (Council of Europe 2001). In other words, what is described must point to learners' or users' accomplishments rather than their weaknesses. Caution should also be observed in using hedging and degree adverbs, such as "comparatively" and "in general," for scaling purposes. Long and complex-structure descriptors were also revised, since CSE users may have found themselves at a loss as to what should be focused on in an individual descriptor. Ambiguity, vagueness, atypical language activities, and linguistic jargon should be avoided wherever possible.

Second, the intended construct of an individual descriptor should be unique. This may be particularly true for descriptors of the translation and interpretation subscales. In some cases, there may be more than one focus in the descriptor, which may give rise to potential misunderstanding. Third, each descriptor follows a three-element model (Pearson Standards and Quality Office 2014): *performance*, *criteria*, and *conditions*. For example, there are three elements in the following descriptor: *Can briefly retell the story with the help of a teacher. Retell the story* is the performance element; *briefly* serves as the criterion, describing how well learners can retell the story; and *with the help of a teacher* is an indicator of the conditions or the prerequisite for the can-do statement. As such, the elements of performance and

criteria, which stipulate the "doing" of the English language and the degree of achievement, were compulsory for all the descriptors. The condition element was optional, given its role of adding or removing constraints. All the descriptors were screened and revised based on the above guidelines. After this process, the descriptor pool was reduced to about 5000 descriptors.

Validating the CSE

In order to report on the research methods, this part describes how the CSE was validated in the two field studies of the last phase of this project.

The Participants

In the first field study, which was larger in scale, the CSE developers decided on how to sample English teachers, learners, and users based on representativeness. Taking into account cross-regional economic development discrepancies, the demographic backgrounds of the participants, as well as the differences in teaching quality across different educational stages in China, the CSE developers finally settled on stratified sampling, where the total numbers of participants from different provinces, municipalities, or autonomous regions in China were first determined. Then participant numbers were allocated to different educational stages, with a specific aim of striking a balance between urban and suburban/rural areas. In accordance with the working levels mentioned above, the participants included Grade 3 and Grade 6 primary school teachers and learners, Junior 3 and Senior 3 high school teachers and learners, college English teachers and learners, and English majors and their supervisors. To facilitate the representation of other stakeholders in the CSE, the working groups also included teachers and students at polytechnics or community colleges and human resources staff from some established enterprises in China.

Approximately 130,000 learners and nearly 30,000 English teachers and related professionals responded to the questionnaires (described below). The working group members were drawn from almost every corner of China. The number of participating schools and universities reached about 1500.

The Questionnaires

As described above, a total of nearly 5000 descriptors, including descriptors of listening, reading, speaking, writing, translation, interpretation, organizational knowledge, and pragmatic knowledge, were pooled together. As it was impractical to request participants to respond to one questionnaire containing such a large number of items, the CSE developers, largely based on the working levels, split the descriptors into 80 questionnaires with about 70 items (descriptors) in each. In order to equate responses from the same proficiency level across different questionnaires in the data analysis, about 20% of the anchor items were spread across the 80 questionnaires so that the responses could be compared. Most of the questionnaires were administered via a self-developed website (http://cse.neea.edu.cn) to gain the

participants' written consent and also to facilitate data collection and follow-up analysis. In some less developed regions, paper questionnaires were delivered and collected for data input.

All the questionnaires used a 5-point Likert scale: 0, 1, 2, 3, and 4 (North 2000). The teacher participants were asked to assign a score to one of their students who characterized the intermediate level of that particular working group. The learner participants were asked to self-rate their own performance against each descriptor. At the ends of the scale, 0 means the student cannot perform what a descriptor says under any circumstances, while 4 means she/he can do so in any conditions. A score of 1 represents a marginal pass, where favorable conditions, such as teacher facilitation, might be required to assist performance; in comparison, a score of 3 means satisfactory performance, where the student can perform despite certain unfavorable conditions such as noise. A score of 2 is in the middle, indicating that the student can do what is described in normal situations.

To ensure the validity of the data and a high return rate, the participants, especially the teacher participants, attended on-site workshops. The working group members introduced the background of the CSE project and clearly stated its purposes and significance. At the workshops the participants were able to familiarize themselves with how and why they should assign different scores to each item (descriptor). Those who could not attend the workshops in person had the option to attend a webinar, where similar instructions were delivered online.

The Data Analysis

The last phase was primarily for scaling the descriptors. More specifically, the cutoff points for different levels were determined. The CSE developers first screened the descriptive statistics and the data distribution to find out if there was any "noise," such as invalid responses and certain outlying cases. Then the participants' responses were standardized based on the predetermined anchor items, so that their judgments on the descriptors could be quantified in the form of logit values and further scaled along a continuum, where difficulty levels as reflected by IRT logit were compared. The more positive the IRT logit value, the more difficult a descriptor seems to be (see Liu 2015a, b for more detail). By doing this, the researchers could investigate the extent to which the working levels could be justified, and adjacent levels could be statistically distinguished.

With these working levels, the CSE developers tried and compared three possibilities: Model 1, a symmetrical scaling with equal intervals; Model 2, a symmetrical scaling with unequal intervals; and Model 3, an asymmetrical scaling with unequal intervals. The most ideal model should meet two requirements. First, the scaling results should be able to accommodate the largest possible number of existing descriptors. However, if the scaling was solely dependent on IRT difficulty values, this could mean sacrificing a number of existing descriptors. Therefore, there must be rounds of adjustment to optimize both IRT values and the existing descriptors. Second, the scaling results should be interpretable to a maximum extent. The different levels should be aligned to certain educational phases to enhance reader-friendliness.

Table 2 Scaling the descriptors

Levels	N	Mean	Range
CSE 1	136	−1.8088	~−2.39
CSE 2	287	−1.5043	−2.39 to −1.65 (0.74)
CSE 3	439	−0.9443	−1.65 to −0.95 (0.70)
CSE 4	510	−0.2726	−0.95 to −0.27 (0.68)
CSE 5	501	0.0643	−0.27 to 0.40 (0.67)
CSE 6	691	0.4464	0.40 to 1.08 (0.68)
CSE 7	658	0.6901	1.08 to 1.78 (0.70)
CSE 8	529	1.3525	1.78 to 2.52 (0.74)
CSE 9	213	1.7202	2.52~

To reach a compromise, the CSE developers proposed a model that resembles Model 1, namely, symmetrical scaling with equal intervals. However, it should be noted that the intervals between adjacent levels are not necessarily absolutely equal. Table 2 lists the scaling results after the first field study. As shown, the zero logit occurs at the CSE 5 level. The IRT difficulty values spread the descriptors along a continuum, with about 0.7 logit (0.7 ± 0.04) as an interval. This method of scaling not only ensured comparatively equal intervals between adjacent levels but also accommodated the existing descriptors to the best possible extent.

Methodologically, how the descriptors were scaled was also similar to the practice of the CEFR developers, who used 1 logit as the absolute and equal interval (North 2000). In the case of the CSE, 0.7 logit was used for three reasons. First, the CSE tended to be more spread out than the CEFR, as there were nine working levels for the CSE. Given its nature of being more fine-grained, this narrowing of logits for scaling intervals may be justifiably accepted. Second, it may be seen in Table 2 that 0.7 logit ensures the symmetry of the scales. The IRT difficulty value range (0.74) between CSE 1 and CSE 2 is the same as that between CSE 8 and CSE 9. Very similar ranges can also be found between CSE 2–CSE 3 and CSE 7–CSE 8 (0.70) and between CSE 3–CSE 4 and CSE 6–CSE 7 (0.68). In addition, as CSE 4, CSE 5, and CSE 6 accounted for the largest number of English learners in China, whose proficiency levels may not be clearly demarcated, the smaller logit intervals can also be justified. In these levels a larger number of descriptors were retained for clearer demarcation. Third, it was found that taking 0.7 logit as the interval ensured that adjacent levels could be distinguished to the greatest extent. If a wider interval (>0.70) was taken, the borderline cases (descriptors) might be pooled together. Likewise, if a narrower interval (<0.70) were used, the number of descriptors for certain levels would have been scant.

The Second Field Study

In the first field study, participants at a particular working level responded to corresponding questionnaires, which facilitated horizontal standardization. However, it did not allow vertical scaling. Therefore, another round of testing was conducted. In this design, a special test was constructed which consisted of all the anchor items at all levels and some existing items which showed good fit,

representing content across all the levels (Kolen and Brennan 2014). The scaling test was controlled to be of a length that could be administered in a single sitting (about 80 test items). Respondents from all of the nine levels were asked to complete the same scaling test. Because some items were too difficult for respondents at the elementary level, special instructions were given to the respondents informing them that there would be difficult items on the test, and that they could respond "0" if a test item was beyond their abilities.

Data from the scaling test was used to construct the score scale. Each respondent taking the scaling test also took the test designed for his or her level. These data were used to link scores at each test level to the scaling test. In addition, as the CSE was intended to be not only descriptive but also prescriptive, the descriptors were further revised based on the results of the first field study. This was also aimed at offsetting the weakness of an overreliance on quantitative data (Liu and Peng 2017). Bearing these concerns in mind, the CSE developers conducted the second field study with participants at different educational stages from six provinces and municipalities in China, allowing for a fair representativeness of demographic distribution.

Therefore, the purposes of the second field study were twofold: cross-validation of the grading and vertical scaling. The CSE developers collected 80 high-quality descriptors that fit the proposed model in the first field study, and then grouped them into a universal questionnaire, which was administered to participants at different educational phases. Next, based on the revised descriptors provided by the eight working groups, more questionnaires were constructed. However, unlike the questionnaire to address the first issue, the new questionnaires were intended to facilitate further culling or revision of the descriptors.

Methodologically, though the second field study still used questionnaires, it was in fact contextualized in an interview setting. At each interview the interviewer read aloud each descriptor in the questionnaires to elicit judgments (scores) and comments from ten interviewees. Two or three working group members took down the interviewees' comments on the spot and also digitalized the scores that the interviewees assigned to each descriptor. The juxtaposition of the interview data with the questionnaire data created the possibility to remove certain descriptors that were judged poorly.

The results from the second field study gave further support to the scaling results from the first field study. The revised descriptors also underwent another round of "re-revision" before the CSE descriptors were finalized. To streamline the understanding of the words used in the CSE, there was also a glossary that defined the terms, such as *familiar topics* and *with ease*. Although most of these were not jargon, they were still explained in detail to avoid possible confusion.

The CSE and English Teaching

As discussed above, the CSE not only serves as standards for English learning and assessment in the Chinese context, but it also provides guidance for EFL teaching. Therefore, this section provides some ideas for how the CSE might be applied in teaching practice.

As companion products of the CSE, there are a number of exemplar language activities which were extracted from the descriptor pool. They shed light on the language activities that could be included in classroom instruction and material compilation at different proficiency levels. From a developmental perspective, certain language activities can rehearse or reinforce learning. Because of space limitations, we take exemplar activities in writing as an example.

From a large pool of written expression descriptors, the CSE developers extracted an inventory of exemplar writing activities across different domains of language use (Pan 2017). As shown in Table 3, in the educational and social domains for English learners and users in China, there are a number of activities ranging from lower proficiency levels to higher ones. At the lower end in the educational domain, the most useful exemplar writing activity is *to copy*, such as *to copy English alphabet and words*. Beginners are not supposed to produce anything in the target language in its written form. However, when it comes to higher proficiency achievers, they are expected to *write a report* or *write a paper/thesis* in educational settings in China. Likewise, in the social domain exemplar, writing activities are mainly related to how learners and users play the roles of different social agents via written communication, such as *to write an email*, *to write a notice*, and *to post an entry online*.

Therefore, in task-based language teaching and learning in the Chinese context, different levels of writing tasks can be designed (see Ellis 2003; Harmer 2001). For instance, an inventory of exemplar writing activities may serve as a directory, under which many related descriptors may be listed. When designing writing tasks in class or compiling materials for English learners, teachers may refer to the descriptors. Table 4 lists a few descriptors associated with the activity *to do creative writing*. At lower proficiency levels (CSE 2 and CSE 3), learners are supposed to write stories based on various prompts. In this case, "story" seems to be a typical text type that learners at these levels tend to encounter. However, at higher proficiency levels (CSE 4 and CSE 7), tasks such as writing drama and composing poems, which are more literarily conventional and demand higher creativity, can be used.

Table 3 Exemplar writing activities in educational and social domains

Domains	Educational	Social
Writing activities	To copy	To write a letter/an email
	To write based on pictorial/graphic prompts	To make a list
	To write a poster/to make a flyer	To write a plan
	To take notes	To write a note
	To write an abstract	To fill out a form
	To write a summary	To write a diary
	To do creative writing	To write a notice
	To write a paper/thesis	To post an entry online
	To write a composition	
	To write a story	
	To write a report	

Table 4 Descriptors about *to do creative writing*

Levels	Descriptors
CSE 2	Can make up a short story based on pictorial prompts
CSE 3	Can make up a story using vocabulary from learning materials
CSE 4	Can write a one-act drama about familiar campus life
CSE 7	Can compose poems with the teacher's guidance

Conclusion

Beginning with a review of existing language proficiency scales and models of language ability, this chapter describes the development and validation of the CSE, an English language proficiency scale in the Chinese context. The CSE developers, by referring to and operationalizing the CLA model, distilled the essences of existing language proficiency scales while offsetting their shortcomings. The de facto scenario of the Chinese context was also taken into consideration, so that the CSE is specifically and appropriately contextualized.

It should be noted that the CSE is not without flaws. Just like the CEFR, it also needs to be refined and updated so that it can better serve the purposes of teaching, learning, and assessment of the English language education in China, enabling its impact on language education to be visible to relevant audiences in related disciplines and thereby promoting productive cross-fertilization among language proficiency scales on a global basis. At present, monographs on how the CSE should be interpreted and how it can be applied in teaching, learning, and assessment are being published in order to attract a wider readership and gain more public attention.

Cross-References

- ▶ Shifting from Teaching the Subject to Developing Core Competencies Through the Subject: The Revised Senior Middle School English Curriculum Standards (2017 Edition) in China
- ▶ Standardized Language Proficiency Tests in Higher Education
- ▶ Postentry English Language Assessment in Universities

Acknowledgment This chapter was based on the Key Project of Philosophy and Social Sciences "The Development of China's Standards of English" (15JZD049) funded by the Ministry of Education, P. R. China.

References

Alderson JC (ed) (2002) Common European Framework of Reference for Languages: learning, teaching, assessment: case studies. Council of Europe, Strasbourg

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

Alderson JC (2010) The Common European Framework of Reference for Language. Invited seminar at Shanghai Jiao Tong University, Shanghai, China, October, 2010
Alderson JC, Banerjee J (2002) Language testing and assessment (part 2). Lang Teach 35(2):79–113
Alderson JC, Figueras N, Kuiper H, Nold G (2006) Analyzing tests of reading and listening in relation to the Common European Framework of Reference: the experience of the Dutch CEFR construct project. Lang Assess Q 3(1):3–30
Anderson LW, Krathwohl DR (2001) A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Longman, New York
Bachman LF (1990) Fundamental considerations in language testing. Oxford University Press, Oxford
Bachman LF, Palmer AS (1996) Language testing in practice: designing and developing useful language tests. Oxford University Press, Oxford
Bachman LF, Palmer AS (2010) Language testing in practice: developing language assessments and justifying their use in the real world. Oxford University Press, Oxford
Canale M (1983) From communicative competence to communicative language pedagogy. In: Richards JC, Schmidt RW (eds) Language and communication. Longman, London, pp 2–27
Canale M, Swain M (1980) Theoretical bases of communicative approaches to second language teaching and testing. Appl Linguist 1(1):1–47
Carroll JB (1961) The nature of data, or how to choose a correlation coefficient. Psychometrika 35(4):347–372
Carroll JB (1968) The psychology of language testing. In: Davies A (ed) Language testing symposium: a psycholinguistic perspective. Oxford University Press, London, pp 46–69
Celce-Murcia M, Dörnyei Z, Thurrell S (1995) Communicative competence: a pedagogical motivated model with content specifications. Issues Appl Linguist 6(2):5–35
Center for Canadian Language Benchmarks (2012) Canadian language benchmarks 2000: English as a second language for adults. Center for Canadian Language Benchmarks, Ottawa
Centre for Canadian Language Benchmarks (2015) Canadian language benchmarks 2000: theoretical framework. Center for Canadian Language Benchmarks, Ottawa
Council of Europe (2001) The Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press, Cambridge, UK
Council of Europe (2018) Common European Framework of Reference for Languages: learning, teaching, assessment. Companion volume with new descriptors. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989
Cumming A (2009) Language assessment in education: tests, curricula and teaching. Annu Rev Appl Linguist 29:90–100
Dai W, Zhang X (2001) On theoretical enrichment of English language teaching in China. Foreign Lang Res 68(2):1–4
Douglas D (2000) Assessing languages for specific purposes. Cambridge University Press, Cambridge, UK
Ellis R (2003) Task-based language learning and teaching. Oxford Applied Linguistics, Oxford
Figueras N (2012) The impact of the CEFR. ELT J 66(4):477–485
Fulcher G (2004) Deluded by artifices? The common European framework and harmonization. Lang Assess Q 1(4):253–266
Halliday MAK (1973) Explorations in the functions of language. Edward Arnold, London
Halliday MAK (1976) The form of a functional grammar. In: Kress G (ed) Halliday: system and function in language. Oxford University Press, Oxford, pp 101–135
Harmer J (2001) The practice of English language teaching, 3rd edn. Pearson Education, Essex
Hulstijn JH (2007) The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. Mod Lang J 91(4):663–667
Hulstijn JH (2011) Language proficiency in native and nonnative speakers: an agenda for research and suggestions for second-language assessment. Lang Assess Q 8(3):229–249
Hymes DH (1972) On communicative competence. In: Pride J, Holmes J (eds) Sociolinguistics. Penguin, Harmondsworth, pp 269–293
Hymes DH (1973) Toward linguistic competence. Texas working papers in sociolinguistics, Working paper No. 16. Centre for Intercultural Studies in Communication, and Department of Anthropology. University of Texas, Austin

Hymes DH (1982) Toward linguistic competence. Graduate School of Education, University of Pennsylvania, Philadelphia
Jackson H, Stockwell P (2011) An introduction to the nature and functions of language, 2nd edn. Continuum International Publishing Group, London
Kolen MJ, Brennan RL (2014) Test equating, scaling, and linking: methods and practices, 3rd edn. Springer, New York
Lado R (1961) Language testing. McGraw-Hill, New York
Little D (2006) The Common European Framework of Reference for languages: content, purpose, origin, reception and impact. Lang Teach 39(3):167–190
Liu J (2015a) Some thoughts on developing China's common framework for English language proficiency. China Examinations 15(1):7–11
Liu J (2015b) Standards-based foreign language assessment. Foreign Lang Teach Res 47(3):417–425
Liu J, Han B (2018) Theoretical considerations in developing the use-oriented China's standards of English. Mod Foreign Lang 41(1):78–90
Liu J, Peng C (2017) Developing scientific China's standards of English. Foreign Lang World 179(2):2–9
Norris JM (2005) Book review: Common European Framework of Reference for Languages: learning, teaching, assessment. Lang Test 22(3):399–405
North B (2000) The development of a common framework scale of language proficiency. Peter Lang Publishing, New York
North B, Panthier J (2016) Updating the CEFR descriptors: the context. Camb Engl Res Notes 63:16–24
Pan M (2017) Towards exemplary writing activities for the China's standards of English: a systemic-functional-linguistics text typology perspective. Foreign Lang World 179(2):17–24
Pearson Standards and Quality Office (2014) Writing descriptors: guidelines and best practice. Pearson Publishing Ltd, London
Purpura J (2008) Assessing communicative language ability: models and their components. In: Broeder P, Martyniuk W (eds) Language education in Europe: the common European framework of reference. Springer, London, pp 2198–2213
Savignon SJ (1983) Communicative competence: theory and classroom practice; texts and contexts in second language learning. Addison-Wesley, Reading
Spolsky B (2008) Language testing at 25: maturity and responsibility? Lang Test 25(3):297–305
van Ek JA (1975) The threshold level in a European unit/credit system for modern language learning by adults. Council of Europe, Strasbourg
Weir CJ (2005a) Limitations of the Common European Framework of Reference for Languages (CEFR) for developing comparable examinations and tests. Lang Test 22(3):281–300
Weir CJ (2005b) Language testing and validation: an evidence-based approach. Palgrave Macmillan, Basingstoke
Wylie E, Ingram DE (2010) International second language proficiency ratings: general proficiency version for English. International Second Language Proficiency Ratings, Queensland
Yang H (2015) Some thoughts on developing a national foreign language testing and assessment system in China. China Examinations 2015(1):12–15
Yang H, Gui S (2007) On establishing a unified Asian level framework of English language proficiency. Foreign Lang China 4(2):34–37
Zou S, Kong J, Zhang W (2015) On the research status and application prospects of CEFR in China. Foreign Lang China 12(5):24–31