BRITISH COUNCIL

TeachingEnglish

ELT Research Papers 18.06

Investigating the applicability of the Common European Framework of Reference (CEFR) for self-assessment in tertiary writing instruction in China: accessibility, effectiveness, feasibility and usefulness

Huahui Zhao, Beibei Zhao and Huiming Ding

UNIVERSITY OF LEEDS

ELT Research Papers 18.06

Investigating the applicability of the Common European Framework of Reference (CEFR) for self-assessment in tertiary writing instruction in China: accessibility, effectiveness, feasibility and usefulness

Huahui Zhao, Beibei Zhao and Huiming Ding

ISBN 978-0-86355-930-3

10 Spring Gardens
London SW1A 2BN, UK

www.britishcouncil.org

About the authors

Dr Huahui Zhao is a Lecturer in Language Education at the University of Leeds. She obtained her PhD from the University of Bristol and worked as a postdoctoral research fellow in distance education at Umeå University, Sweden, before moving to Leeds. Her research interest lies in the interplay between language assessment and technology-enhanced learning. She has published widely in key journals including *Assessing Writing, Learning and Instruction, British Journal of Education Technology* and *ELT Journal*.

Dr Beibei Zhao obtained her EdD in TESOL/Applied Linguistics from the University of Bristol. She is a Lecturer in English Education at Zhejiang Shuren University, China, and is a co-ordinator of the College English Department. Her primary interests lie in the areas of second language acquisition, language testing/assessment, English for Academic Purposes (EAP) and English language teaching.

Huiming Ding is a doctoral researcher at the University of Leeds. Her research interest focuses on educational assessment, especially the development, use and impact of large-scale assessments. She worked with PISA (Programme for International Student Assessment) data in the China PISA National Centre before starting a PhD, and has a master's degree in Language Testing.

Abstract

The current project investigated the accessibility, feasibility, effectiveness and usefulness of the Common European Framework of Reference for Languages (CEFR) and its companion European Language Portfolio (ELP) descriptors for tertiary writing instruction in China to explore how the CEFR could be used to bridge teaching, learning and assessment.

The project was conducted in three phases, involving two writing tutors and 146 students from three subjects in a university in China. In the pre-assessment phase, participants' perceptions of the original ELP descriptors were sought and used as the basis for modifying the descriptors for self- and teacher assessment. In the assessment phase, learners' self-assessment ratings were compared with teacher assessment ratings. In the post-assessment phase, participants' perceptions of the accessibility, feasibility and usefulness were explored.

The qualitative analysis of pre-assessment perception data revealed learners' difficulties in understanding the original ELP descriptors and therefore bilingual versions with a reduced number of technical words were developed. The quantitative analyses revealed that more than half of the students gave the same ratings as their tutors. However, students tended to assess their writing proficiency at a significantly higher level than their teachers did, suggesting students' over-estimation of their writing proficiency. The post-assessment survey suggested that students and tutors highly valued the roles of the self-assessment descriptors in raising their awareness of writing weaknesses, setting learning objectives for next assignments and developing their understanding of effective writing. The project substantiated the importance of eliciting learner voices and contextualising the CEFR in local contexts. It also provides important implications for using China's Standards of English Language Ability for teaching and learning.

Contents

1

Introduction

This project explored the applicability of the Common European Framework of Reference for Languages (CEFR) for tertiary English writing in China, in the context of current English testing reform in China, which aims to develop a standard of English language ability that bridges teaching, learning, testing and learner autonomy. The project created self-assessment grids based on the European Language Portfolio (ELP) descriptors to help learners evaluate their own writing, identify their next learning objectives and develop learner autonomy.

Teacher-driven writing instruction prior to the research project

Chinese education is well known for its entrenched teacher-driven learning (Zhao, 2018). The instruction of writing in this research context is no exception. Prior to the introduction of self-assessment into the context, writing instruction was observed to be mainly driven by the writing tutors.

For example, in a lesson on summary writing, the tutors asked students to read a short text and analyse its main ideas and structure. This was followed by a task that asked students to use linking words between sentences and then a task to practise paraphrasing sentences. The tutors then talked about how to use the ideas and structures of the reading text when writing summaries and highlighted the language that could be used to structure summaries. The students then produced their summaries.

Similarly, in a lesson on writing argumentative essays, students were asked to read a short text and then the tutors used the text to explain how to provide supporting details to main ideas, how to write topic sentences and how to draw a conclusion. This was followed by practice tasks for the passive voice and linking words, respectively. The tutors then used the textbook for the writing course to teach students how to use a process-oriented writing approach before they started to write their argumentative essays on a similar topic to that of the reading text.

Assessment of both writing genres was conducted solely by the writing tutors who used a rather narrow marking band (i.e. 65–75 out of 100) to indicate the quality of student writing. Because of the relatively large class sizes, little information was provided to justify marks and explain the strengths and weaknesses of student writing. No follow-up activities were carried out, which resulted in the production of only one draft, despite the important role of revisions in the process-approach to writing.

China’s Standards of English Language Ability (CSE)

The Chinese government, along with English educators in China, have been aware of the drawbacks of existing education practice, in particular its limitations in developing learners’ English language ability/communicative competence and learner autonomy. To address these limitations, the State Council of the People’s Republic of China launched the development of a new standard of English language ability in 2014 to strengthen the relationship between teaching, learning and testing, and improve consistency between the test scores and test takers’ language proficiency, and the variety of local, regional and national English tests (Jiang, 2016). Based on Jiang (2016), the scale would heavily draw upon the CEFR to develop descriptors for learning, teaching and assessment while taking into consideration the specific Chinese English education context.

In 2018 after the launch of the current project, China’s Ministry of Education and State Language Commission published *China’s Standards of English Language Ability* (hereafter referred to as the CSE), which describes nine levels of English language ability based on ‘can do’ statements (China’s Ministry of Education and State Language Commission, 2018). The CSE consists of quantitative scales and qualitative descriptors to define different levels of listening, speaking, reading, writing and translation/interpretation ability as well as other aspects of language such as grammar, phonology, lexis, sentence structure, organisation, genre and register. No guidance has been provided on how to use the standards in teaching, learning and assessment. There is now the need for research to explore the application of the new standards for language teaching.

2

The overall aim of the project

This project was a pilot project to explore how 'can do' statements could be used to bridge teaching, learning and assessment in the Chinese educational context. It developed self-assessment scales for writing tasks, using ELP descriptors as the basis. In particular, the project explored the applicability of the existing ELP descriptors by eliciting teachers' and learners' perspectives on its accessibility. The respondents' views were used to modify the ELP descriptors, and the effectiveness of using the modified descriptors for self-assessment was evaluated using teacher assessment of the same descriptors as the comparison baseline. The study also explored the usefulness and feasibility of using the CEFR and ELP descriptors in the Chinese tertiary writing context. The project website can be found here: https://cefrinchina.leeds.ac.uk/

The CEFR and ELP

Given their importance in the current research, we will briefly introduce the CEFR and ELP descriptors before discussing the research design.

Brief introduction of the CEFR

The CEFR was created and published by the Council of Europe in 2001 to provide 'a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe' (Council of Europe, 2001a: 1). The framework provides three scales: the global scale, the self-assessment grid and the illustrative scale (Council of Europe, 2001a, 2001b). The self-assessment grid comprises a set of 'can do' statements to provide language learners with a checklist to reflect on their language learning stages and plan their learning objectives. Learners using the grid select the one description which best approximates their current level of English proficiency. As such, the time to complete the self-assessment task is relatively short (Association of Language Testers in Europe, 2002).

The self-assessment grid is, however, rather broad and the 'can do' statements do not address specific assessment focuses related to writing tasks. For example, self-assessment for writing at the B2 level includes the statement: 'I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view.' This long statement needs to be unpacked into more detailed aspects of language ability for students to accurately assess what they can actually do. Chinese English tutors in Zheng and her colleagues' (2016) study also reported that the condensed information in such descriptors made it difficult to understand the self-assessment grid. In addition, some B1 statements are about genres or topics that are not always the focus of writing instruction in English for Academic Purposes (EAP), such as personal letters in the descriptor: 'I can write personal letters describing experiences and impressions' and topics in: 'I can write straightforward topics, which are familiar, or of personal interest'.

Brief introduction of the ELP descriptors

Considering the aforementioned limitations of the CEFR self-assessment grid, the European Language Portfolio (ELP) descriptors were used as the basis to create self-assessment scales in the current project. The ELP descriptors were also developed by the Council of Europe in parallel with the CEFR (Lenz and Schneider, 2004). Compared with the CEFR descriptors, the ELP descriptors are more specific, tapping into different aspects of language use in different contexts (Lenz and Schneider, 2004). For example, the portfolio provides an additional nine descriptors to assess overall written production across the six proficiency levels of the CEFR. The ELP descriptors also expand the CEFR descriptors for reports and essays, covering more writing aspects including, but not limited to, linguistic accuracy, structure, content, genres and familiarity with topics.

The ELP descriptors were designed for different levels of language learners including tertiary-level adult students. The primary objective of the ELP descriptors is to encourage learners to 'reflect on their language learning, set targets, record progress and document their skills. They are an effective aid to developing independence and a capacity of self-directed learning, and so are useful in language study' (Council of Europe, 2001a: 1). This objective aligns with the purpose of the current project: to support the development of autonomous language learners through the use of self-assessment that requires learners to reflect on what they already know and how well they know it, in order to help them plan the next stage of learning. Therefore, the ELP descriptors were used for the self-assessment of EAP writing in the project, a demanding but essential language skill for English language learners in China.

3

Research design

To examine the applicability of the ELP descriptors for EAP writing in higher education in China, four research questions were asked:

1. What are students' and teachers' perceptions of the ELP descriptors in terms of accessibility and usefulness?
2. Are modifications to the ELP descriptors necessary to make them applicable for Chinese English learners and, if so, what should these be?
3. To what extent do learners' self-assessment ratings agree with those of teacher assessment using the modified ELP descriptors?
4. What are students' and teachers' perceptions of the feasibility and usefulness of the modified ELP descriptors for self-assessment of EAP writing?

The first question investigated the applicability of the pre-modified ELP descriptors from participants' perspectives through a survey with students after they used them in a training session alongside tutors' reflective logs (Q1). The responses provided the basis for the modifications of the ELP descriptors for the later formal assessment phase (Q2). After discussion and consultation with the writing tutors, the agreed modified versions of the ELP descriptors were used for self- and teacher assessment and ratings from the two groups were compared to evaluate the effectiveness of self-assessment using the modified ELP descriptors (Q3). The fourth question explored the potential usefulness of the ELP descriptors in self-assessment for the development of students' EAP writing from students' perspectives via a survey and teachers' perspectives via reflective logs.

Participants

Two tutors and 146 second-year students from four classes and three subjects at the Zhejiang Shuren University (ZSU), China, participated in the project for one semester on a voluntary basis. Both tutors and student participants were Chinese and spoke English as a foreign language. One teacher participant held a master's degree in English Literature (Tutor 1 teaching Classes 1 and 2) while the other held a doctoral degree in Language Education (Tutor 2 teaching Classes 3 and 4). Both had been working at the ZSU for more than ten years and had taught the academic reading and writing integrated module since its introduction in 2016.

The student participants were second-year university students who were majors of Network Media (Classes 1 and 2), Public Management (Class 3), and Chinese Linguistics and Literature (Class 4). Most of the students had been learning English for more than ten years since primary school, with an approximate English proficiency level around B1–B2 judging by their entrance English exam scores, their writing scores at the end of the academic year and their tutors' assessment. They had very limited experience in self-assessment owing to the traditional teacher-driven and examination-oriented learning culture in China.

Table 1 summarises the participants' background information. As shown in Table 1, Classes 1 and 2 were from the same subject group, and consisted of a more balanced number of male and female students than Classes 3 and 4. Descriptive analysis and an ANOVA (analysis of variance) test of English writing proficiency across classes showed no significant difference in writing proficiency among Classes 1 to 3 but students in Class 4 had a significant 5 point higher average writing score than the rest of the three classes. Gender and writing proficiency were found to significantly affect the agreement of self- and teacher assessment, but discussion of these variables is beyond the scope of the current study.

Table 1: Student participants' background

Class ID	Number	Male	Female	Subject	Final writing scores
1	35 (taught by Tutor 1)	16	19	Network Media	69.69 (SD=7.66)
2	35 (taught by Tutor 1)	13	22	Network Media	67.79 (SD=7.17)
3	29 (taught by Tutor 2)	8	21	Public Management	71.31 (SD=7.79)
4	47 (taught by Tutor 2)	3	44	Chinese Linguistics and Literature	75.98 (SD=6.22)

The EAP module was introduced in 2016 as an optional module for all students in the host institution. As a result, class composition changed, with smaller class sizes and higher student motivation. Relatively high motivation also made teachers and learners willing to try out new teaching methods including self-assessment to improve teaching and learning. However, considering participants' limited experience of self-assessment, training was provided in the first writing session wherein the original ELP descriptors were introduced to the writing class.

Research phases

The project was carried out in three phases. The pre-assessment phase was designed to elicit the students' and tutors' perceptions of the original descriptors after training in self-assessment with the ELP descriptors. Their perception data was used to modify the descriptors. The modified descriptors were then used in the assessment phase where students and teachers conducted self- and teacher assessment. In the post-assessment phase, the participants' experience of using the modified ELP descriptors in self-assessment was investigated.

Phase 1: pre-assessment

In the pre-assessment phase, students were asked to use the original ELP descriptors to assess their summary assignments and subsequently filled in a survey to report their perceptions of the accessibility and usefulness of the descriptors in self-assessment.

Creation of pre-modified ELPs

The limited number of studies investigating the CEFR in the Chinese education context has suggested mixed results in terms of its applicability. Zou and Zhang (2017) reported on adapting can do statements in the Chinese higher education context, and Zheng et al. (2016) reported on tutors' difficulty in understanding them. The limited use of the ELP descriptors in the Chinese context means it is difficult to evaluate the accessibility and feasibility of the descriptors for any particular group of learners. Therefore, it was felt necessary to investigate the target participants' perceptions of the descriptors and to make modifications based on those perceptions, if necessary.

To create the self-assessment grid for the training session, the ELP descriptors were selected and collated in line with the lesson plans and participating students' existing English proficiency levels. Descriptors of reporting and essays from A2 to C1 with the majority from B1 and B2 were selected from the ELP bank and then mixed in the pre-modified self-assessment grids. The two genres aligned approximately with the two writing tasks that were addressed in the module, namely: summaries and argumentative essays. Descriptors at the four levels of English language proficiency were included to address individual differences and were presented in mixed order in the grids. The descriptors aimed to prompt students to self-assess both macro- and micro-writing skills in terms of structuring summaries/argumentative essays and the language used in them, respectively. For instance, the pre-modified descriptors of summaries consisted of 20 items, the first nine items focusing on structuring summaries and the remaining 11 items on the language used in summaries. Three emoticons were used for students to assess their writing proficiency: ☺ standing for achieved, 😐 standing for nearly there and ☹ standing for not there yet.

This pre-modified version for summaries was then piloted among the participating students in the training session. Given the limited class time, the pre-modified version for argumentative essays was not pilot tested. However, students' perceptions of the self-assessment grid of summaries were used to modify the self-assessment grids for argumentative essays.

Training in self-assessment

Training in self-assessment using ELP descriptors was provided to address learners' lack of knowledge of how to assess themselves (Little, 2009). A one-hour training session was provided to students in four stages:

1. discussing the advantages of self-assessment, learners' concerns over self-assessment and how to carry out self-assessment
2. introducing the pre-modified version of ELP descriptors and their use and popularity in language learning
3. tutors demonstrating how they used the ELP descriptors to assess students' writing
4. students reading and then using the ELP descriptors to assess a summary.

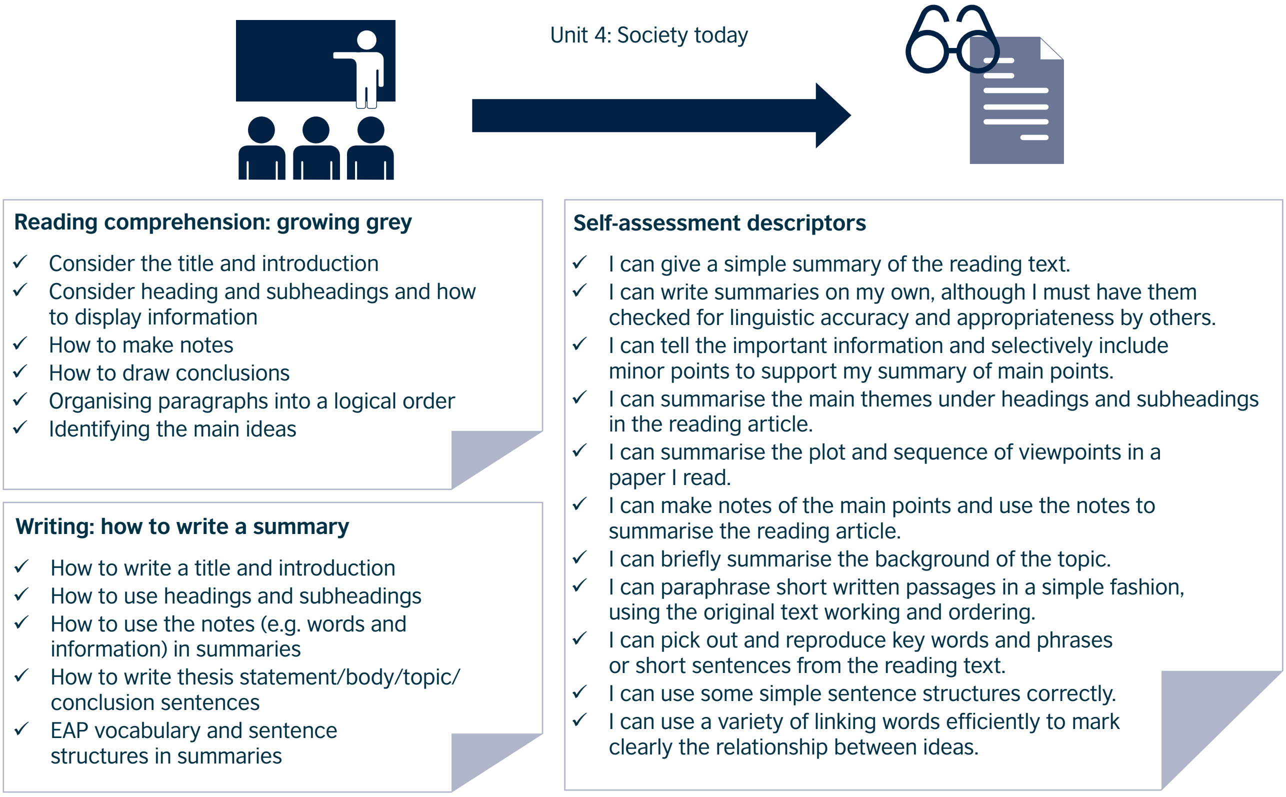
The training session concluded with students' reflections on the ELP descriptors in terms of their accessibility (i.e. how well they could understand the ELP descriptors), usefulness and reasons behind their responses when completing the pre-assessment questionnaire (referred to as PRE hereafter) (Appendix A). The students' reflections provided the empirical evidence for the modifications of the ELP descriptors that were used at the assessment stage.

Phase 2: assessment

In the assessment phase, students were asked to self-assess four essays, comprising two summaries and two argumentative essays. The four essays focused on four topics, including society today/growing grey, food security/the challenge of feeding nine billion people, sustainable energy/using waste and sustainable fashion. Students wrote summaries of the first two topics based on their reading of relevant texts and argumentative essays on the last two topics and their associated reading texts. After they had written their essays, they were given 30 minutes to conduct self-assessment and to reflect on their writing experience by ticking one of the three options in the self-assessment grid.

Each 'can do' statement in the self-assessment grid focused on a key point from the content of the lesson. As such, self-assessment served as a tool for learners to reflect on what was taught in each lesson and on whether they had achieved the learning objectives, and then for learners to plan their next stage of learning activities. Figure 1 depicts the relationship between the focus of Unit 4 and the related self-assessment activity.

Figure 1: Alignment between teaching, learning and assessment



Reading comprehension: growing grey

✓ Consider the title and introduction
✓ Consider heading and subheadings and how to display information
✓ How to make notes
✓ How to draw conclusions
✓ Organising paragraphs into a logical order
✓ Identifying the main ideas

Writing: how to write a summary

✓ How to write a title and introduction
✓ How to use headings and subheadings
✓ How to use the notes (e.g. words and information) in summaries
✓ How to write thesis statement/body/topic/conclusion sentences
✓ EAP vocabulary and sentence structures in summaries

Self-assessment descriptors

✓ I can give a simple summary of the reading text.
✓ I can write summaries on my own, although I must have them checked for linguistic accuracy and appropriateness by others.
✓ I can tell the important information and selectively include minor points to support my summary of main points.
✓ I can summarise the main themes under headings and subheadings in the reading article.
✓ I can summarise the plot and sequence of viewpoints in a paper I read.
✓ I can make notes of the main points and use the notes to summarise the reading article.
✓ I can briefly summarise the background of the topic.
✓ I can paraphrase short written passages in a simple fashion, using the original text working and ordering.
✓ I can pick out and reproduce key words and phrases or short sentences from the reading text.
✓ I can use some simple sentence structures correctly.
✓ I can use a variety of linking words efficiently to mark clearly the relationship between ideas.

Teacher support was available during the whole self-assessment session. Self-assessment grids were collected after each assessment session and stored in a folder separate from teacher assessment.

Teacher assessment of the same essays was conducted using the same ‘can do’ statements outside class sessions. Teachers did not read the self-assessment ratings in order to avoid the possible influence of the self-assessments on teacher assessment. Therefore, for each assessment task, there were two document folders for self- and teacher assessment ratings. The two sets of data were compared to evaluate the effectiveness and reliability of self-assessment. Although it is arguable that teacher assessment is not necessarily reliable, we would suggest that in a classroom context, the tutors are the main assessment agents who measure the quality of student writing, and respond with potential adjustments of teaching and learning approaches. Therefore, agreement between teacher and self-assessment indicates a shared understanding of student writing proficiency within a class, which in turn will enable the adoption of appropriate strategies needed by student writers to support learning.

Phase 3: post-assessment

After the final self-assessment session, the participating students were asked to fill in a short survey to express their views on the accessibility, feasibility and usefulness of the modified ELP descriptors in self-assessment. Comparable questions were asked to the two writing tutors based on their observation of students’ use of the descriptors and their own experience of using the descriptors. Instead of using interviews, the two writing tutors were provided with prompts to write their reflections on the use of the modified ELP descriptors in self- and teacher assessment. One advantage of reflective logs over interviews is that they support longer and deeper thought over the questions than face-to-face interviews. Another advantage of reflective logs is flexibility in time and space. This was particularly important for the project as the two tutors needed to prepare students for final examinations at the end of the project and they were heavily involved in related administration duties, in addition to their teaching. The flexibility of the reflective logs supported the teachers in participating in the study and in providing richer data.

4

Key research findings

In this section, findings are reported in terms of the accessibility of pre-modified ELP descriptors from the participants' perspectives, the modification of the descriptors, the agreement between teacher and self-assessment, and the feasibility and usefulness of the modified descriptors in self-assessment.

Accessibility of pre-modified ELP descriptors in self-assessment

The accessibility of pre-modified ELP descriptors was investigated from the learners' and tutors' perspectives. The perception results were used as the basis for modifying the descriptors to accommodate the target learners' needs.

Students were asked to indicate their understanding of ELP descriptors on a four-point scale, from 1 = extremely easy to 4 = extremely difficult (PREQ4), after they had used the descriptors to assess their own summaries in the training session. Table 2 suggests that all the participants thought the ELP descriptors were of a relatively low accessibility level, with Class 1 finding them easier than the other three classes and Class 2 finding them more difficult than the other classes, despite Classes 1 and 2 sharing the same subject. This could be because Class 2 had a slightly lower level of English writing proficiency (mean = 67.79, SD = 7.17) than Class 1 (mean = 69.69, SD = 7.66).

When the students were asked to identify the specific items that they felt difficult (PREQ5), the items in Table 3 were mentioned most frequently.

Table 2: Accessibility of pre-modified ELP descriptors

Class ID	N	Mean	Standard deviation (SD)
1	33	2.30	.64
2	35	2.91	1.20
3	28	2.68	.67
4	47	2.62	.491
Total	143	2.63	.80

Table 3: Difficult items that were mentioned more than ten times

Difficult items	Frequencies
Item 11	25
Item 23	25
Item 10	23
Item 21	19
Item 5	17
Item 18	13
Item 9	11
Item 4	10

We can observe that Items 11 and 23 were the most difficult items for students, followed by Items 10 and 21. Qualitative analysis via the qualitative data analysis software package NVivo 11 suggested that among the 37 responses, 19 responses claimed *unknown words* as the main reason for difficulty in understanding the ELP descriptors. Statements such as 'I don't know some of words in this item' and 'I can't understand the meaning of these difficult words' were reiterated in responses. Fifteen responses indicated low English language proficiency including 'poor' grammatical knowledge, inability to express themselves clearly in English and low English language proficiency in general as the main reasons. Fifteen responses claimed that the difference between English and Chinese made it hard to understand the descriptors: 12 referred to the different sentence structure while three referred to different language use. Seven responses suggested that it was hard to understand the focus of the ELP descriptors because of limited understanding of academic writing.

Students' perceptions of the accessibility of the self-assessment grid were echoed by the two teachers in their reflection logs. Tutor 2 was more optimistic than Tutor 1 and believed that most of her students should be capable of understanding the ELP descriptors well, although she thought the descriptors could be difficult for a small number of learners with low English proficiency. By contrast, Tutor 1 expected that most of her students would find it difficult to understand the descriptors considering their low English language proficiency. She further commented that similar items placed close to each other would increase the difficulty level, such as Item 2 ('I can summarise the short text by using words from the original reading text') and Item 3 ('I could find keywords, phrases and short sentences in the original reading materials and use them to summarise the short text').

The learners' and teachers' views on the difficult wording and sentence structures are in line with previous findings by Gori (2011) in his study on adapting the ELP descriptors in Italy. He highlighted the teacher-oriented nature of the CEFR and difficulties that were caused by the high level of formality and the high frequency of technical words. To integrate participant voice in the modification of the descriptors in the present study, students' and teachers' suggestions for revising the descriptors were elicited (PREQ6). Unsurprisingly, using simpler and less technical language was the most frequent advice given to improve the accessibility of descriptors, followed by the suggestions for using Chinese in self-assessment grids.

Modifications of ELP descriptors

In response to participants' suggestions, the most important modification was reducing the technical terms and formality of the language by creating a bilingual version of descriptors for at least three reasons. First, six students mentioned the benefits of reading the ELP descriptors in English (e.g. learning new vocabulary and being aware of their limited vocabulary sizes). Therefore, it was thought beneficial to retain the English descriptors because this exposure to the target language in learning would promote learner autonomy (Little, 2009). Second, a bilingual version would be more effective in solving the difficulties caused by the differences between English and Chinese language than simply replacing the technical words with easier vocabulary. This, as the two teachers suggested, could result in the loss of meaning in the descriptors. Finally, although a bilingual version could distract learners from the descriptors in English, considering the main aim of the self-assessment activity was to encourage learners to reflect on their learning progress, the accessibility of the descriptors should be prioritised so that learners could reflect on their learning progress accurately. Bearing in mind the three reasons above, a bilingual version seemed to be a fair compromise to remedy the difficulties of understanding the ELP descriptors.

Another way to make the descriptors more accessible was to decrease the assessment items for each session and thus reduce learners' cognitive load and allow learners more time to reflect on their writing performance. Therefore, descriptors in the self-assessment grid of summaries in the training session were divided into two parts across two consecutive writing sessions. One session focused on self-evaluation of structuring a summary (Appendix B) while the other session encouraged students to reflect on their language use in summaries (Appendix C). Table 4 provides an overview of the focus of each self-assessment session throughout the research period.

Table 4: Focus of each self-assessment session

Self-assessment session	Assessment focuses
Session 1	Training in self-assessment using pre-modified ELP descriptors
Session 2	Constructing summaries (9 modified items)
Session 3	Language use in summaries (12 modified items)
Session 4	Constructing argumentative essays (7 modified items)
Session 5	Language use in argumentative essays (14 modified items)

In addition, the tutor's comments on neighbouring items with similar focus causing confusion was addressed by closely examining descriptors and relocating similar ones further apart. Although the pre-assessment questionnaire responses were based on self-assessment descriptors for summaries, the modifications were also applied to the self-assessment descriptors for the two argumentative tasks as well as the teacher assessment grids, which were created by changing 'I can do' to 'she/he can do' statements. This supported the evaluation of the agreement between self- and teacher assessment as an indication of the effectiveness/reliability of self-assessment.

Effectiveness of the ELP descriptors for self-assessment: agreement between self- and teacher assessment

The effectiveness of the ELP descriptors for supporting self-assessment was evaluated using teacher assessment ratings on the same assignments as the comparison baseline. The agreement between self- and teacher assessment ratings was carried out on a three-point scale based on the three emoticons that were used for self- and teacher assessment: 1 = ☺ achieved, 2 = 😐 nearly there and 3 = ☹ not there yet. In other words, a lower rating (i.e. a smaller number) stood for a higher achievement. Inter-rater agreement tests and Wilcoxon signed-rank tests were carried out to examine the agreement between self- and teacher assessment ratings for each descriptor on the same piece of writing. The results were reported in the order of the four tasks, namely: Summary 1 (structuring a summary), Summary 2 (the language use in a summary), Argument 1 (structuring an argumentative essay) and Argument 2 (the language use in an argumentative essay).

Agreement between self- and teacher assessment ratings in Summary 1

Cohen's Kappa was used to analyse the agreement between self- and teacher assessment ratings of students' competence in structuring summaries.

Table 5: Kappa inter-rater reliability between self- and teacher assessment scores in Summary 1

Descriptors	Kappa value	Asymptotic standard error[a]	Approximate T[b]	Approximate significance
D1	-.008	.063	-.148	.883
D2	.146	.076	2.069	.039*
D3	.241	.072	3.564	.000*
D4	.208	.068	3.710	.000*
D5	.389	.073	5.432	.000*
D6	.039	.063	.650	.516
D7	.166	.062	3.343	.001*
D8	.159	.054	3.875	.000*
D9	-.001	.051	-.027	.978
N of valid cases	134			
[a] Not assuming the null hypothesis.				
[b] Using the asymptotic standard error assuming the null hypothesis.				
* Statistically significant results.				

Table 5 showed significant agreement in six out of the nine descriptors (p<0.05), suggesting that students and teachers reached the same judgement about these six aspects of structuring summaries across a significant number of assignments. However, the small Kappa values for the six descriptors with a range of 0.001 and 0.389 suggested a relatively low level of agreement in the whole data set. In other words, only a small number of students gave the same rates as the tutors. This indicated the need for a test of difference for self- and teacher assessment scores. Wilcoxon signed-rank tests were conducted to observe the difference between self- and teacher assessment ratings owing to the skewed distribution of the data. Table 6 shows significant differences existing in seven out of the nine assessment descriptors (p<0.05).

Table 6: Differences between self- and teacher assessment ratings in Summary 1

	TAS1D1 – SAS1D1	TAS1D2 – SAS1D2	TAS1D3 – SAS1D3	TAS1D4 – SAS1D4	TAS1D5 – SAS1D5	TAS1D6 – SAS1D6	TAS1D7 – SAS1D7	TAS1D8 – SAS1D8	TAS1D9 – SAS1D9
Z	-5.253[a]	-.692[a]	-1.838[a]	-3.095[a]	-3.130[a]	-4.866[b]	-5.416[a]	-5.336[a]	-3.252[a]
Sig. (2-tailed)	.000*	.489	.066	.002*	.002*	.000*	.000*	.000*	.001*
[a] Based on negative ranks.									
[b] Based on positive ranks.									
* Statistically significant results.									
Note: TAS1 = teacher assessment in Summary 1; SAS1 = self-assessment in Summary 1.									

Table 7 provides further information regarding similarities or differences between self- and teacher assessment for each descriptor. Key points to observe in Table 7 include:

- More than half of the assignments (i.e. equal to or more than 67 out of 134 assignments in total) had the same score from teachers and students for six out of the nine descriptors (i.e. ties), which were Descriptors 2–5 and 7–8. This further confirms the significant agreements that were reported based on Cohen's Kappa tests above.
- More assignments received a higher (i.e. positive rank, lower achievement) than lower (i.e. negative rank, higher achievement) score from teachers than from the students for all the descriptors except Descriptor 6. The result suggests that students might over-estimate their proficiency in structuring their summaries in these aspects compared with their teachers or the teachers might be harsher than the students in assessing these aspects.

Table 7: Ranks of teacher assessment and self-assessment ratings in Summary 1

		N	Mean rank	Sum of ranks
TAS1D1 – SAS1D1	Negative ranks[a]	17	40.00	680.00
	Positive ranks[b]	64	41.27	2641.00
	Ties[c]	53		
	Total	134		
TAS1D2 – SAS1D2	Negative ranks	29	33.74	978.50
	Positive ranks	36	32.40	1166.50
	Ties	69		
	Total	134		
TAS1D3 – SAS1D3	Negative ranks	22	29.50	649.00
	Positive ranks	36	29.50	1062.00
	Ties	76		
	Total	134		
TAS1D4 – SAS1D4	Negative ranks	16	28.75	460.00
	Positive ranks	40	28.40	1136.00
	Ties	78		
	Total	134		
TAS1D5 – SAS1D5	Negative ranks	11	21.00	231.00
	Positive ranks	31	21.68	672.00
	Ties	92		
	Total	134		
TAS1D6 – SAS1D6	Negative ranks	61	40.83	2490.50
	Positive ranks	18	37.19	669.50
	Ties	55		
	Total	134		
TAS1D7 – SAS1D7	Negative ranks	8	28.00	224.00
	Positive ranks	49	29.16	1429.00
	Ties	75		
	Total	132		
TAS1D8 – SAS1D8	Negative ranks	11	32.50	357.50
	Positive ranks	54	33.10	1787.50
	Ties	69		
	Total	134		
TAS1D9 – SAS1D9	Negative ranks	32	42.88	1372.00
	Positive ranks	60	48.43	2906.00
	Ties	42		
	Total	134		

[a] Negative ranks: teacher assessment ratings are lower than self-assessment ratings.

[b] Positive ranks: teacher assessment ratings are higher than self-assessment ratings.

[c] Ties: teacher assessment ratings are equal to self-assessment ratings.

Agreement between self- and teacher assessment ratings in Summary 2

Cohen's Kappa was used to analyse the agreement between self- and teacher assessment of the language use in summaries (Summary 2).

Table 8: Inter-rater reliability between teacher and self-assessment scores in Summary 2

Descriptors	Number of valid case	Kappa value	Asymptotic standard errora	Approximate T[b]	Approximate significance
D1	131	.099	.075	1.321	.186
D2	130	.255	.067	4.498	.000*
D3	131	.161	.056	2.759	.006*
D4	129	.052	.081	.652	.514
D5	131	.098	.048	2.108	.044*
D6	130	.004	.069	.065	.948
D7	131	.130	.063	2.282	.022*
D8	123	.173	.067	3.197	.001 *
D9	132	–.028	.065	–.437	.662
D10	130	.142	.056	2.58	.010*
D11	131	.248	.066	4.062	.000*
D12	131	.286	.068	4.389	.000*
[a] Not assuming the null hypothesis.					
[b] Using the asymptotic standard error assuming the null hypothesis.					
* Significant Kappa value.					

Table 8 shows significant agreement existing in eight of the 12 descriptors (p<0.05), suggesting students and teachers provided the same ratings on these eight aspects of language use in summaries across a significant number of assignments. However, the small Kappa value for each descriptor with a range of 0.004 and 0.286 suggests a relatively low level of agreement in the whole data set. This indicates the need for a test of difference for the self- and teacher assessment scores.

Table 9 shows that Wilcoxon signed-rank tests revealed significant differences in ten out of the 12 assessment descriptors (p<0.05). Table 10 provides further information on self- and teacher ratings for each descriptor. Similar to Summary 1, more than half of the assignments received the same rating (i.e. ties) from teachers and students for all the descriptors except Descriptor 10. Moreover, more assignments received a higher rating (i.e. positive rank, lower achievement) from teachers than from students for all the descriptors except Descriptors 2 and 4. The results indicate that students might over-estimate their proficiency in their language use in summaries.

Table 9: The differences between teacher assessment and self-assessment ratings in Summary 2

	TAS2D1 – SAS2D1	TAS2D2 – SAS2D2	TAS2D3 – SAS2D3	TAS2D4 – SAS2D4	TAS2D5 – SAS2D5	TAS2D6 – SAS2D6	TAS2D7 – SAS2D7	TAS2D8 – SAS2D8	TAS2D9 – SAS2D9	TAS2D11 – SAS2D10	TAS2D10 – SAS2D11	TAS2D12 – SAS2D12
Z	-1.156[a]	-2.248[b]	-5.724[a]	-.938[b]	-6.575[a]	-2.281[a]	-2.609[a]	-1.983[a]	-5.063[a]	-4.400[a]	-5.632[a]	-3.459[a]
Sig. (2-tailed)	.248	.025*	.000*	.348	.000*	.023*	.009*	.047*	.000*	.000*	.000*	.001*
[a] Based on negative ranks.												
[b] Based on positive ranks.												
* Statistically significant differences.												

Table 10: Ranks of teacher and self-assessment ratings in Summary 2

		N	Mean rank	Sum of ranks
TAS2D1 – SAS2D1	Negative ranks[a]	26	31.88	829.00
	Positive ranks[b]	36	31.22	1124.00
	Ties[c]	69		
	Total	131		
TAS2D2 – SAS2D2	Negative ranks	34	26.79	911.00
	Positive ranks	18	25.94	467.00
	Ties	78		
	Total	130		
TAS2D3 – SAS2D3	Negative ranks	7	33.86	237.00
	Positive ranks	54	30.63	1654.00
	Ties	70		
	Total	131		
TAS2D4 – SAS2D4	Negative ranks	31	29.11	902.50
	Positive ranks	25	27.74	693.50
	Ties	73		
	Total	129		
TAS2D5 – SAS2D5	Negative ranks	10	35.50	355.00
	Positive ranks	69	40.65	2805.00
	Ties	52		
	Total	131		

		N	Mean rank	Sum of ranks
TAS2D6 – SAS2D6	Negative ranks	26	34.35	893.00
	Positive ranks	44	36.18	1592.00
	Ties	60		
	Total	130		
TAS2D7 – SAS2D7	Negative ranks	17	28.74	488.50
	Positive ranks	38	27.67	1051.50
	Ties	76		
	Total	131		
TAS2D8 – SAS2D8	Negative ranks	17	24.41	415.00
	Positive ranks	31	24.55	761.00
	Ties	75		
	Total	123		
TAS2D9 – SAS2D9	Negative ranks	16	38.00	608.00
	Positive ranks	60	38.63	2318.00
	Ties	56		
	Total	132		
TAS2D10 – SAS2D10	Negative ranks	15	28.00	420.00
	Positive ranks	48	33.25	1596.00
	Ties	67		
	Total	130		
TAS2D11 – SAS2D11	Negative ranks	6	24.50	147.00
	Positive ranks	48	27.88	1338.00
	Ties	77		
	Total	131		
TAS2D12 – SAS2D12	Negative ranks	14	25.00	350.00
	Positive ranks	38	27.05	1028.00
	Ties	79		
	Total	131		

[a] Negative ranks: teacher assessment ratings are lower than self-assessment ratings.

[b] Positive ranks: teacher assessment ratings are higher than self-assessment ratings.

[c] Ties: teacher assessment ratings are equal to self-assessment ratings.

Agreement between self- and teacher assessment ratings in Argument 1

Table 11 shows significant agreement in five of the seven descriptors ($p<0.05$), suggesting students and teachers provided the same ratings for the five descriptors across a significant number of assignments. However, the small Kappa value for each descriptor with a range of 0.050 and 0.366 suggests a low level of agreement in the whole data set. This indicates the need for a test of difference between self- and teacher assessment ratings.

Table 11: Inter-rater reliability between self- and teacher assessment scores in Argument 1

Descriptors	Measure of agreement: Kappa value	Asymptotic standard error [a]	Approximate T [b]	Approximate significance
D1	.366	.073	4.657	.000*
D2	.143	.075	1.885	.059
D3	.050	.060	.849	.396
D4	.212	.068	3.664	.000*
D5	.246	.075	4.759	.000*
D6	.162	.066	2.953	.003*
D7	.162	.067	3.327	.001*
N of valid cases	142			
[a] Not assuming the null hypothesis.				
[b] Using the asymptotic standard error assuming the null hypothesis.				
* statistically significant results.				

Table 12: Differences between teacher assessment and self-assessment ratings in Argument 1

	TAA1D1 – SAA1D1	TAA1D2 – SAA1D2	TAA1D3 – SAA1D3	TAA1D4 – SAA1D4	TAA1D5 – SAA1D5	TAA1D6 – SAA1D6	TAA1D7 – SAA1D7
Z	-2.654[a]	-1.054[b]	-4.336[a]	-4.185[a]	-4.523[a]	-5.333[a]	-3.052[a]
Asymp. sig. (2-tailed)	.008*	.292	.000*	.348	.000*	.000*	.002*
[a] Based on negative ranks.							
[b] Based on positive ranks.							
* Statistically significant results.							

Table 12 shows that Wilcoxon signed-rank tests revealed significant difference in five of the seven descriptors ($p<0.05$). Table 13 further suggests that more than half of the assignments had the same score (i.e. ties) from teachers and students for all the descriptors except Descriptor 3. In addition, more than half of the assignments received a higher rating (i.e. positive rank, lower achievement) from teachers than from students for all the descriptors except Descriptor 1. More than 90 of the 142 assignments received the same rating from teachers and the student writers themselves for Descriptors 1, 5 and 9. The results are similar to those for the two summary tasks and suggest that students may over-estimate their writing competence.

Table 13: Ranks of teacher and self-assessment ratings in Argument 1

		N	Mean rank	Sum of ranks
TAA2D1 – SAA2D1	Negative ranks[a]	14	23.50	329.00
	Positive ranks[b]	32	23.50	752.00
	Ties[c]	96		
	Total	142		
TAA2D2 – SAA2D2	Negative ranks	36	32.78	1180.00
	Positive ranks	28	32.14	900.00
	Ties	78		
	Total	142		
TAA2D3 – SAA2D3	Negative ranks	24	36.75	882.00
	Positive ranks	60	44.80	2688.00
	Ties	58		
	Total	142		
TAA2D4 – SAA2D4	Negative ranks	15	29.50	442.50
	Positive ranks	47	32.14	1510.50
	Ties	80		
	Total	142		
TAA2D5 – SAA2D5	Negative ranks	7	22.50	157.50
	Positive ranks	37	22.50	832.50
	Ties	98		
	Total	142		
TAA2D6 – SAA2D6	Negative ranks	9	28.00	252.00
	Positive ranks	50	30.36	1518.00
	Ties	83		
	Total	142		
TAA2D7 – SAA2D7	Negative ranks	13	22.50	292.50
	Positive ranks	33	23.89	788.50
	Ties	96		
	Total	142		
[a] Negative ranks: teacher assessment ratings are lower than self-assessment ratings.				
[b] Positive ranks: teacher assessment ratings are higher than self-assessment ratings.				
[c] Ties: teacher assessment ratings are equal to self-assessment ratings.				

Agreement between self- and teacher assessment ratings in Argument 2

Table 14 shows significant agreement in eight of the 14 descriptors (p<0.05). However, the small Kappa value for each descriptor with a range of 0.05 and 0.247 suggests a relatively low level of agreement within the whole data set. This indicates the need for a test of difference between self- and teacher assessment ratings.

Table 15 shows significant difference in 12 of the 14 descriptors (p<0.05).

Table 14: Inter-rater reliability between self- and teacher assessment scores in Argument 2

Descriptors	N of valid case	Measure of agreement: Kappa value	Asymptotic standard error[a]	Approximate T[b]	Approximate significance
D1	140	.194	.049	3.643	.000*
D2	140	.137	.067	2.300	.021*
D3	140	.129	.073	1.868	.062
D4	138	.127	.070	1.834	.067
D5	140	.075	.059	1.291	.197
D6	140	.133	.052	2.648	.008*
D7	140	.141	.067	2.123	.034*
D8	139	.084	.061	1.517	.129
D9	139	.198	.074	3.448	.001*
D10	140	.247	.057	4.506	.000*
D11	140	.119	.068	1.880	.060
D12	140	.055	.053	1.174	.240
D13	138	.109	.055	2.398	.017*
D14	140	.245	.067	3.862	.000*

[a] Not assuming the null hypothesis.

[b] Using the asymptotic standard error assuming the null hypothesis.

* Statistically significant differences.

Table 15: Differences between teacher assessment and self-assessment ratings in Argument 2

	TAA2D1 – SAA 2D1	TAA2D2 – SAA 2D2	TAA2D3 – SAA 2D3	TAA2D4 – SAA 2D4	TAA2D5 – SAA 2D5	TAA2D6 – SAA 2D6	TAA2D7 – SAA 2D7	TAA2D8 – SAA 2D8	TAA2D9 – SAA 2D9	TAA2D10 – SAA 2D10	TAA2D11 – SAA 2D11	TAA2D12 – SAA 2D12	TAA2D13 – SAA 2D13	TAA2D14 – SAA 2D14
Z	-7.437[a]	-4.418[a]	-1.485[a]	-4.299[b]	-5.894[a]	-6.745[a]	-1.641[a]	-3.389[a]	-3.960[a]	-5.728[a]	-2.693[a]	-6.099[a]	-5.968[a]	-3.501[a]
Sig. (2-tailed)	.000*	.000*	.138	.000*	.000*	.000*	.101	.001*	.000*	.000*	.007*	.000*	.000*	.000*

[a] Based on negative ranks.

[b] Based on positive ranks.

* Statistically significant differences.

Table 16 suggests that more than half of the assignments have the same ratings from teachers and students for all the descriptors except Descriptors 5 and 6 (i.e. ties). Descriptor 9 is the most salient descriptor and was given the same rank by 89 out of the 139 students and by the tutor. In addition, all assignments received a higher rating (i.e. positive rank, lower achievement) from teachers than from students for all the descriptors except Descriptor 4. Far more assignments received higher ratings (low achievement) from the tutors than the student writers themselves for Descriptors 1, 5, 6, 10, 12 and 13. The results echo the findings of the previous three tasks and suggest that students may over-estimate their writing proficiency and appropriate language use in argumentative essays.

Table 16: Ranks of teacher and self-assessment ratings in Argument 2

		N	Mean rank	Sum of ranks
TAA2D01 – SAA2D1	Negative ranks[a]	3	33.50	100.50
	Positive ranks[b]	65	34.55	2245.50
	Ties[c]	72		
	Total	140		
TAA2D02 – SAA2D2	Negative ranks	15	29.50	442.50
	Positive ranks	49	33.42	1637.50
	Ties	76		
	Total	140		
TAA2D03 – SAA2D3	Negative ranks	27	32.20	869.50
	Positive ranks	38	33.57	1275.50
	Ties	75		
	Total	140		
TAA2D04 – SAA2D4	Negative ranks	52	34.77	1808.00
	Positive ranks	16	33.63	538.00
	Ties	70		
	Total	138		
TAA2D05 – SAA2D5	Negative ranks	13	38.00	494.00
	Positive ranks	65	39.80	2587.00
	Ties	62		
	Total	140		
TAA2D06 – SAA2D6	Negative ranks	8	29.00	232.00
	Positive ranks	68	39.62	2694.00
	Ties	64		
	Total	140		
TAA2D07 – SAA2D7	Negative ranks	27	34.52	932.00
	Positive ranks	41	34.49	1414.00
	Ties	72		
	Total	140		
TAA2D08 – SAA2D8	Negative ranks	19	34.03	646.50
	Positive ranks	48	33.99	1631.50
	Ties	72		
	Total	139		

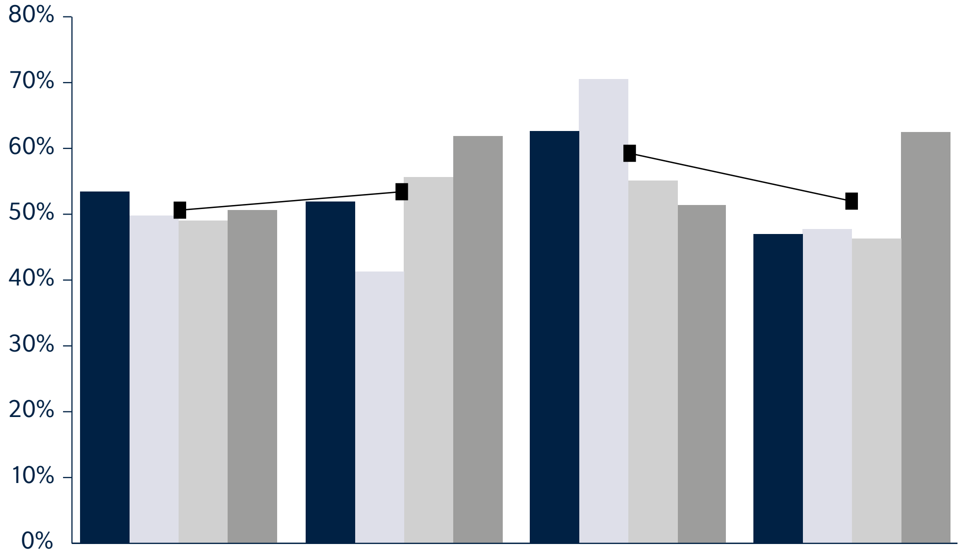
		N	Mean rank	Sum of ranks
TAA2D09 – SAA2D9	Negative ranks	11	25.50	280.50
	Positive ranks	39	25.50	994.50
	Ties	89		
	Total	139		
TAA2D10 – SAA2D10	Negative ranks	9	30.50	274.50
	Positive ranks	55	32.83	1805.50
	Ties	76		
	Total	140		
TAA2D11 – SAA2D11	Negative ranks	23	33.00	759.00
	Positive ranks	44	34.52	1519.00
	Ties	73		
	Total	140		
TAA2D12 – SAA2D12	Negative ranks	10	33.00	330.00
	Positive ranks	62	37.06	2298.00
	Ties	68		
	Total	140		
TAA2D13 – SAA2D13	Negative ranks	9	33.00	297.00
	Positive ranks	58	34.16	1981.00
	Ties	71		
	Total	138		
TAA2D14 – SAA2D14	Negative ranks	16	28.50	456.00
	Positive ranks	42	29.88	1255.00
	Ties	82		
	Total	140		
a. Negative ranks: teacher assessment ratings are lower than self-assessment ratings.				
b. Positive ranks: teacher assessment ratings are higher than self-assessment ratings.				
c. Ties: teacher assessment ratings are equal to self-assessment ratings.				

Summary of agreement between self- and teacher assessment across the four tasks

The inter-rater reliability analysis along with difference analysis reveals consistent findings across the four tasks: significant and relatively low agreement, and significant differences between student writers and writing tutors for both macro- and micro-aspects of writing summaries and argumentative essays. Descriptive analysis showed more than half the assignments received the same ratings from the students and the writing tutors for most of the descriptors.

To further explore the effectiveness of the ELPs for self-assessment, percentage agreement analyses between teacher and self-assessment were carried out. Figure 2 confirms the previous findings as it reveals that, on average, more than half the assignments received the same ratings from teachers and the students in the four tasks.

Figure 2: Agreement between self- and teacher assessment across tasks and classes



	SUM1	SUM2	ARG1	ARG2
■ Agreement in Class 1	53.41%	51.95%	62.61%	46.96%
□ Agreement in Class 2	49.79%	41.28%	70.56%	47.77%
░ Agreement in Class 3	49.04%	55.68%	55.10%	46.31%
▒ Agreement in Class 4	50.09%	61.89%	51.37%	62.46%
■—■ Average % of agreement	50.62%	53.42%	59.26%	51.99%

However, there are slight discrepancies in the agreement percentages among tasks and classes. On average, the task of how to structure an argument (Argument 1) achieved the highest agreement percentage among the four tasks. Class 4 received the highest average agreement percentage among the four classes (56.45 per cent), followed by Class 1 (53.73 per cent). Class 3 had the lowest average agreement between self- and teacher assessment (51.53 per cent). Undoubtedly, the differences across classes would also be influenced by various other factors including the genre of tasks, the writing tutors and the student background. This suggests the need to analyse factors that could potentially impact learners' self-assessment behaviour and consequently affect the agreement between self- and teacher assessment.

Students' perceptions of the usefulness of ELP-based self-assessment for writing

Students were asked about the usefulness of the ELP descriptors for the development of their English writing proficiency prior to and after using the modified ELP descriptors (PREQ7 and POSTQ1). Of the students, 96.9 per cent and 97.9 per cent perceived the activities to be useful based on the pre-modified and post-modified descriptors, respectively.

Furthermore, the number of students who considered the self-assessment activities as 'very useful' or 'extremely useful' in the post-assessment survey increased by 9.5 per cent compared with the pre-assessment survey. An additional paired t-test showed the difference between pre- and post-assessment findings regarding the usefulness of self-assessment was not significant ($p = 0.07$); however, based on a scale of 1–4 (1 = extremely useful and 4 = not useful), the mean decreased from 2.72 (SD = 0.65) to 2.61 (SD = 0.65), suggesting increasing perceived usefulness of self-assessment activities for writing proficiency.

The quantitative results were supported by the students' self-reports on the usefulness of self-assessment for their English writing development. Sixty-eight of the 117 (58 per cent) students in the survey (PREQ6) after the training session stated that the activity was useful for them to understand their writing proficiency, particularly in terms of identifying those aspects of writing that needed improvement. Similarly, 73 of the 135 (54 per cent) students reported in the post-assessment questionnaire that the self-assessment activities were useful in terms of identifying their weaknesses in writing. Of those 73 students, 22 further explained that the ELP-based self-assessment activities made them reflect on their writing weaknesses and address them in their next assignments (e.g. improving clarity of their writing and the use of linking words). A number of students suggested that the self-assessment activities improved their motivation as they knew what they needed to do in their next assignments. A number of students thought that the self-assessment activities developed their ability to self-evaluate their own learning progress.

On the other hand, 12 students suggested that the self-assessment activities were not useful because although the ELP descriptors could help them become aware of their weaknesses, they did not know how to improve those areas. Statements such as 'It can make me sometimes find my own problems. But I don't know how to fix it' and 'It was not that useful to improve my writing proficiency even though it helped me to be aware of my weak areas' were repeated. Therefore, they suggested that teachers should integrate follow-up activities to address those weak areas. A few students further suggested including advice after those descriptors that were rated as *nearly there* and *not there yet* to help them develop those areas of writing.

The two writing tutors echoed the students' statements in the post-assessment reflective log. Both believed that the self-assessment activities had helped their students become more aware of their writing weaknesses and those areas they should work on. In addition, they thought that the self-assessment descriptors made learners think of various other aspects of writing apart from grammar and vocabulary and thus developed their understanding of what a good piece of writing should be. This might be a greater benefit than ratings of self-assessment. As for the tutors themselves, the teacher assessment descriptors developed their understanding of essential facets of good summaries and argumentative essays, which helped them to decide on instruction focus in class, including accuracy, structure, coherence and cohesion as indicated in descriptors. Furthermore, through the self-assessment activities alongside the teacher assessment, they realised that students were capable of assessing themselves. One tutor believed 80 per cent of the students could arrive at similar ratings to teacher assessment based on their classroom observation. They planned to continue to use similar self-assessment activities in their future teaching.

Students' perceptions of the feasibility of ELP-based self-assessment for writing

Students were asked about the feasibility of using the modified ELP descriptors in self-assessment activities after the four self-assessment sessions (POSTQ5–6). A mean of 2.92 (SD = 0.72) was achieved, which suggested a moderate to high feasibility level for the descriptors for self-assessment activities based on a scale of 1–5 (1 = extremely easy and 5 = extremely difficult). In addition, among the 243 respondents, 23.8 per cent (i.e. 34) students thought that the grid was extremely easy or easy for them, whereas 15.4 per cent students considered that it was difficult, including three students who thought the ELPs were extremely difficult to use. The moderate feasibility suggested by the mean was confirmed by 60.8 per cent of students holding a neutral position towards the feasibility of the ELPs in self-assessment activities.

When the students were asked to explain their responses on the low feasibility for the descriptors (POSTQ7), four main reasons emerged:

1. Fifteen students stated that lack of experience of self-assessment resulted in their uncertainty of how to assess their own English proficiency despite the training session and the modified descriptors.
2. Twenty-one students believed their difficulty in understanding the descriptors affected how they felt about the feasibility of using the descriptors. Their claims were supported by the significant association between the feasibility (POSTQ5) and the accessibility of the modified grid (POSTQ3) ($F_{1, 141}$ = 51.10, p<0.001). A variance of 26.6 per cent in perceived feasibility could be explained by the accessibility of the descriptors for the students.

3. Students reported their uncertainty about their own language proficiency, which led to the difficulty of rating themselves using the modified ELP descriptors. Students said their fluctuating feelings about their own writing proficiency on different occasions made them unsure whether they had assessed themselves accurately. A few students explained that uncertainty about their own language proficiency made them hesitant in giving an accurate rating of the descriptors. Similarly, students felt that the feasibility of the ELPs was affected by their limited language proficiency, which prevented them from fully understanding the descriptors and how they could be used to assess their writing. Among them, a few students admitted that those descriptors asking for lower levels of writing aspects were more helpful (e.g. word usage) than those for higher levels of aspects (e.g. coherence) because they could act only on the former but not the latter.
4. Students thought the assessment grids did not include a complete checklist for writing, which led to their difficulty in using them. Some students felt that some of the descriptors seemed to be rather similar. Other students felt the descriptors were too broad and needed more detail.

The tutors believed the feasibility for the modified descriptors was very high based on their classroom observation as fewer students asked about the descriptors and students seemed to be more confident and efficient in rating themselves. However, one tutor believed the feasibility of the ELPs for self-assessment could be further improved by adding more detail to some of the descriptors to help students understand them. Nevertheless, both tutors believed that the modified ELP descriptors were feasible, although the students needed to be more committed to the self-assessment activities, echoing a few students' admission to their lack of commitment to the self-assessment activities in their responses to POSTQ7. Both tutors also explained that some students were not very committed to the self-assessment activities, possibly because of their low motivation for learning English.

Both students' and tutors' perceptions of the feasibility of using the ELP descriptors suggested possible further modifications could be adopted to improve their feasibility and accessibility. Further qualitative analysis of each descriptor would be helpful to identify those descriptors needing further revisions.

Summary and discussion

The current project explored the accessibility, feasibility, reliability/effectiveness and usefulness of the CEFR in general and the ELP descriptors in particular for self-assessment of EAP writing in China's tertiary education. The pre-assessment survey on the accessibility of the ELP descriptors revealed a relatively low level for accessibility and learners' difficulties in understanding the descriptors mainly owing to technical words and the formality of the language, and the students' developing English language proficiency. Drawing on students' and tutors' suggestions, the descriptors were modified to create a bilingual version with fewer descriptors and reduced cognitive load for each task. Using the modified descriptors, students assessed their own writing proficiency in structuring summaries and argumentative essays and using appropriate language and writing conventions in these two genres. The tutors provided teacher assessment using the same descriptors outside classes.

The results showed that more than half of the students could assess their writing competence reliably, using teacher assessment ratings as the comparison baseline. Meanwhile, variation in the agreement percentages of self- and teacher assessment was observed across tasks and classes. The students and tutors reported an intermediate to relatively high level of the feasibility for the modified descriptors. They also expressed overwhelmingly positive views on the usefulness of the self-assessment activities, particularly in terms of raising their awareness of the weaknesses of their writing and developing their understanding that good writing involved other aspects in addition to grammar and vocabulary. On the other hand, students reported that limited self-assessment experience and their developing English language proficiency affected their understanding of the self-assessment grids and consequently their accurate assessment of their own proficiency. Further modifications of the descriptors were also suggested by the learners and teachers on how to improve accessibility and feasibility.

The agreement between self- and teacher assessment and the perceived high level of feasibility and usefulness of self-assessment grids have demonstrated the potential for applying the CEFR in higher education in China with appropriate modifications and support for students. On the one hand, the project has generated evidence that learners at B1–B2 levels are able to carry out the CEFR-based self-assessment with appropriate support and accessible and feasible self-assessment grids. On the other hand, the project has revealed potential difficulties in using the CEFR in an educational context outside the European zone and hence the necessity to adapt the CEFR to the context of use. The project has suggested important implications for applying the CEFR and its associated ELP descriptors in local texts.

First, the CEFR and the ELP descriptors are good bases for creating grids for assessment for learning (in this case self- and teacher assessment of writing). This echoes the main aim of the framework and the descriptors as supporting teaching and learning. In this project, the framework has served to build a bridge between predominant summative assessment and neglected formative assessment in local contexts. Prior to the project, writing assessment only involved the provision of a mark with little formative feedback for how to improve writing quality. The current project selected and then modified the original ELP descriptors to encourage learners and tutors to reflect on what student writers have achieved, nearly achieved and not achieved based on the writing product in line with learning outcomes for each teaching session. The results suggest that the CSE, which are heavily based on the CEFR, could be used to guide teaching, learning and assessment in English classrooms with appropriate modification and adaption.

Second, the limitations of the CEFR and the ELP descriptors could discourage their use in teaching and learning. Students expressed their difficulties in understanding the pre-modified descriptors. After modification, students continued to have difficulty in understanding the post-modified descriptors, based on the post-assessment survey. Difficulties in understanding the descriptors could negatively affect learners' and tutors' perceptions of the feasibility and usefulness of the CEFR and their later use of the descriptors in assessment, a position supported by the statistically significant relations between accessibility and feasibility of assessment grids and the agreement between self- and teacher assessment. To make the descriptors more accessible and feasible, technical words could be modified and the information in a descriptor could be explained and unpacked in several descriptors, if necessary. Most importantly, students' voices should be heard and then integrated into the modifications of descriptors. In other words, the students should share if not have the sole ownership of the descriptors so that they can understand what they are assessing and why they should assess themselves in such a way.

Last but not least, the differences among tasks and classes suggest the need to invest effort in designing accessible and feasible self-assessment grids. The effectiveness of self-assessment could be further affected by the complexity of the interwoven relationships of student factors (e.g. gender, subjects, writing proficiency and English proficiency), task factors (e.g. genres and focuses of assessment) and tutor factors (e.g. tutors' cognition about the framework and assessment and teachers' practice of teaching and assessment). The intertwined relationships suggest the importance of learner agency (students being actively involved in the whole process of designing, using and reflecting on self-assessment activities), teacher agency (teachers being committed to designing accessible and feasible assessment grids across tasks and student groups) and the negotiation between learner and teacher agency. The use of new descriptors in self-assessment requires tutors to adapt their roles in designing the self-assessment grids, to provide instruction/support during the use of self-assessment grids and to facilitate rather than decide assessment for developing writing literacy.

Conclusion

The project was launched when the Chinese government initiated the development of the new standards of English language ability in which the CEFR had been heavily influential. The main aim of the project was to explore how the CEFR and therefore the CSE could be used to facilitate learning rather than facilitate the measurement of learning achievement. The findings of the project set out a promising vision for using the CSE to bridge teaching, learning and assessment in the Chinese education context. At the same time, the project revealed the potential challenges of using the CSE, for which possible solutions were provided, including engaging learners' and tutors' voices to create tailored assessment grids, integrating the learning outcomes of teaching sessions into the assessment descriptors and designing follow-up activities to support learners in achieving learning objectives.

It is expected that the current pilot project will raise teachers' awareness of the reciprocal relationships between summative and formative assessment and encourage language educators to embark on an innovative but appropriate use of summative descriptors to facilitate learning through student-driven assessment. The findings of the project have begun to make an impact on teacher cognition and practice through a research-informed workshop held in September where language educators at the University of Leeds were invited to discuss and share their experience of using the CEFR in language teaching (Appendix D). The dissemination of the research report will, we hope, encourage more language educators to consider the formative roles of high-stake summative standards in their teaching and consequently use testing for learning rather than merely as assessments of learning.

References

Association of Language Testers in Europe (2002) The ALTE can do project.

China's Ministry of Education and State Language Commission (2018) China's Standards of English Language Ability.

Council of Europe (2001a) Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of Europe (2001b) Common European Framework of Reference for Languages: Learning, teaching, assessment: Structured overview of all CEFR scales. Cambridge: Cambridge University Press.

Gori, F (2011) The translation of the CercleS ELP into Italian: a case study on the AICLU version. *CercleS 2011* 1/1: 171–177.

Jiang, G (2016) Implementing suggestions on deepening the reform of examination and enrolment system and steadily advancing the national assessment system of foreign language proficiency. *China Examinations*, 3–6.

Lenz, P and Schneider, G (eds) (2004) A bank of descriptors for self-assessment in European Language Portfolios. Council of Europe, Language Policy Division.

Little, D (2009) *The European Language Portfolio: where pedagogy and assessment meet*. Paper presented at the 8th International Seminar on the European Language Portfolio, 29 September – 1 October 2009, European Centre for Modern Languages, Graz, Austria.

Zhao, H (2018) Exploring tertiary English as a Foreign Language writing tutors' perceptions of the appropriateness of peer assessment for writing. *Assessment & Evaluation in Higher Education*, 1–13.

Zheng, Y, Zhang, Y and Yan, Y (2016) *Investigating the practice of the Common European Framework of Reference for Languages (CEFR) outside Europe: a case study on the assessment of writing in English in China*. ELT Research Paper, London: British Council.

Zou, S and Zhang, W (2017) Exploring the adaptability of the CEFR in the construction of a writing ability scale for test for English majors. *Language Testing in Asia* 7/1: 18.

Appendix A:
Survey on pre-modified ELP descriptors

Investigating the applicability of the CEFR for self-assessment in tertiary writing instruction in China: accessibility, feasibility, effectiveness and usefulness

Dear All,

This questionnaire aims to investigate your perceptions of the self-assessment grid. Please read questions, instructions and options carefully. There are no 'right' or 'wrong' answers. As this questionnaire is intended for research purposes only, the information provided is considered anonymous, confidential and will not be disclosed to third parties without your permission.

I truly appreciate your volunteering to co-operate and spend time completing the questionnaire. This questionnaire consists of eight questions. You will need about 10 minutes to complete the questionnaire. Thank you.

Section one: biographic information (your student ID number): □

1. What is your subject? (Please tick the option appropriate for you.)

a. Arts □

b. Science □

c. Engineering □

d. Others □, please specify here: ____________________

2. What is your English exam score in the College Entrance Examination?

3. Have you passed the following exams? (*You may choose more than one option.*)

a. CET-4 □ what is the result? □

b. CET-6 □ what is the result? □

c. TOEFL □ what is the result? □

d. IELTS □ what is the result? □

Section two: your viewpoints of the self-assessment grid

4. What do you think about the accessibility of the assessment grid?

a. Extremely easy to understand □

b. Easy to understand □

c. Difficult to understand □

d. Extremely difficult to understand □

5. Which descriptor or descriptors are difficult for you to understand? Please write down the order number of the descriptor or descriptors (e.g. 1a) and explain why (e.g. difficult wording, not applicable to you, difficult to measure, etc.).

□

6. Can you think of how to improve the descriptor or descriptors that you mentioned above? Please comment on them one by one if you identify more than one descriptor.

7. Is the assessment grid useful for you to assess and understand your writing proficiency?

a. Extremely helpful □

b. Very helpful □

c. Helpful □

d. Not helpful □

8. Please explain your answer to Question 7.

9. Do you have any other comments about the assessment grid?

Thank you for completing the questionnaire.

Appendix B:
Self-assessment grid: Summary (1)

Student Number: ________________________________ Class: ________________________________

Structure a summary	☺	😐	☹
1. I can give a simple summary of the reading text. 我能写简单的文章概述。			
2. I can paraphrase short written passages in a simple fashion, using the original text wording and ordering. 我能用原文的措词和顺序简单地概述短小的文章。			
3. I can pick out and reproduce key words and phrases or short sentences from the reading text. 我能从原文中找到关键词、短语或者短句并用于文章概述。			
4. I can tell the important information from minor one and selectively include minor points to support my summary of main points. 我能区分重要信息和次要信息，并能有选择性地用次要信息来帮助概述重要信息。			
5. I can write summaries on my own, although I must have them checked for linguistic accuracy and appropriateness by others. 我能自如地概述，虽然我需要检查语言的精确性和正确性。			
6. I can summarise the plot and sequence of viewpoints in a paper I read. 我能概述文章的情节与其关联的观点。			
7. I can summarise the main themes under headings and subheadings in the reading article. 我能概述文章中标题和小标题下的主题。			
8. I can make notes of the main points and use the notes to summarise the reading article. 我能记笔记，并用笔记来概述文章。			
9. I can briefly summarise the background of the topic. 我能简要地概述话题的背景知识。			

Appendix C:
Self-assessment grid: Summary (2)

Student Number: ______________________________ Class: ______________________________

Language use in summaries	☺	😐	☹
1. I can have good control of elementary vocabulary, but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations. 我能较好地使用基本的词汇，但是当我表达比较复杂的思想或者处理不熟悉的话题和情形的时候，会出现较大的错误。			
2. I can write with reasonable grammatical accuracy and can correct mistakes if they are identified by others. 我能相对准确地使用语法，并能纠正别人指出的语法错误。			
3. I can write with accurate punctuation. 我能正确地使用标点符号。			
4. I can spell accurately, apart from occasional slips of the pen. 我能正确地拼写单词，除了个别笔误的情况。			
5. I can use some simple sentence structures correctly. 我能正确地使用简单的句子结构。			
6. I can use some simple structures correctly but may mix up tenses and forget to mark agreement. 我能正确地使用一些简单的句子结构，但是会混淆时态和忘记主谓一致。			
7. I can use a variety of linking words efficiently to mark clearly the relationships between ideas. 我能有效地使用各种各样的连接词，清楚地标注各种关系。			
8. I can link a series of shorter, discrete simple elements into a connected, linear sequence of points. 我能把一系列短小的、零散的成分串成有关联的点。			
9. I can use the most frequently occurring connectors to link simple sentences like ‘and’, ‘but’ and ‘because’. 我会用诸如“和”、“但是”、“因为”等使用频率很高的简单的关联词连接句子。			
10. I can make it clear what I am trying to express although language errors could occur. 我能清楚地表达自己的想法，尽管会有语言方面错误发生。			
11. I can qualify opinions and statements precisely in relation to degrees of, for example, certainty/ uncertainty, belief/doubt, likelihood, etc. 我能不同程度地限定观点和陈述，比如确定/不确定，相信/怀疑，类似，等等。			
12. I can convey simple information of immediate relevance, getting across which point I feel is the most important. 我能传达简单并直接关联的信息，表达我认为最重要的信息。			

Appendix D:
ELTRA Workshop flyer

Explore the use of CEFR for language teaching: benefits and challenges

Date: 18 September 2018
Time: 9.30 a.m.–1.30 p.m.
Venue: Hillary Place SR (G.18)
Facilitator: Dr Huahui ZHAO (Lecturer in Language Education, specialized in language assessment and testing)

Abstract: This workshop aims to promote the good practice of using CEFR (Common European Framework for Reference), the most widely used framework for language assessment, in language teaching, drawing upon key findings of a completed project funded by British Council. During this participant-oriented and research-based workshop, you will be invited to design and reflect on the use of CEFR for your targeted students in terms of accessibility, feasibility, benefits and challenges. The workshop will be concluded via sharing implications of the completed project for the use of CEFR in language teaching.

Agenda

Time	Item
9.30 a.m.–9.40 a.m.	Registration and welcome
9.40 a.m.–10.20 a.m.	Introducing CEFR, ELPs, peer and self-assessment in language teaching/learning
10.20 a.m.–11.30 a.m.	Designing peer and self-assessment grids with CEFR descriptors
11.30 a.m.–12.00 p.m.	Reflecting on design experience in terms of benefits and challenges
12.00 p.m.–12.30 p.m.	Lunch break (lunch provided)
12.30 p.m.–1.10 p.m.	Project report and discussion
1.10 p.m.–1.30 p.m.	Wrap up and look ahead

Contact: Please register your attendance and any dietary or special access requirements to Dr Huahui Zhao via emailing h.zhao1@leeds.ac.uk. 20 places are available on a first-come, first-served basis.

ISBN 978-0-86355-930-3