

RESEARCH

Open Access



Exploring the adaptability of the CEFR in the construction of a writing ability scale for test for English majors

Shen Zou^{1*} and Wenxing Zhang²

* Correspondence:

shzou@shisu.edu.cn

¹School of English Studies, Shanghai International Studies University, No.550, Dalian Rd (W), Shanghai 200083, People's Republic of China
Full list of author information is available at the end of the article

Abstract

Background: The CEFR, ever since its inception, has had profound impact on language teaching, learning, and assessment not only in Europe but also in other parts of the world. This study focuses on the adaptability of CEFR writing descriptors in the context of Test for English Majors (TEM).

Methods: First, we constructed a questionnaire based on the descriptors collected from various sources in order to elicit university teachers' views on the importance of these descriptors. A revised version was produced based on the feedback from the initial questionnaire survey. In order to further investigate what level or levels these remaining descriptors would fall into, 35 university teachers of English were invited to complete the revised questionnaire while rating 36 TEM writing scripts.

Results: Band-setting of the descriptors was initially determined on the basis of the questionnaire data, the result of which was the draft scale of writing ability. In order to collect further evidence for our calibration of the descriptors, eight university teachers of English were interviewed. Based on the interview data, some descriptors were fine-tuned before the scale was finalized.

Conclusions: The results have shown that CEFR writing descriptors can be used in the description of the writing ability of TEM candidates, but most of the CEFR descriptors surveyed have had their original level altered in our writing ability scale.

Keywords: CEFR, Writing scale, TEM, Descriptors

Background

Following two draft versions (Council of Europe 1996a, 1996b), the Council of Europe officially published the Common European Framework of Reference for Languages (CEFR) in 2001 (Council of Europe 2001). Up to now, it is available in 38 languages, and is being increasingly consulted and used in educational and non-educational contexts inside Europe (Martyniuk 2012, p. 1). Two surveys were conducted by the Council of Europe in 2005 and 2006 to investigate the use of the CEFR at institutional and national levels respectively (ibid.). The results show that the CEFR had a major impact on language education, especially in the field of curricula/syllabi planning and development (ibid.). The CEFR's impact is not confined to language teaching and learning only; its influence has extended to the

realm of language assessment as well. In order to facilitate the alignment process, the Council of Europe published a document *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Manual: Preliminary Pilot Version* (Council of Europe 2003) followed by a reference supplement in the ensuing year *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe 2004). And again in 2009, the Council of Europe 2009 published *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*, which provides further material on maintaining standards across languages, contexts, and administrations by exploiting teacher judgment and IRT scaling. Language examination authorities and language testers have so far attempted to align examinations to the CEFR or design tests or self-assessment instruments based on the CEFR scale (e.g., DIALANG, see <http://www.dialang.org>; see also Khalifa and Ffrench 2008; Huhta et al. 2002; North 2002).

With the growing influence of the CEFR, its impact has also gone beyond Europe. For example, in Taiwan, a common standard of English proficiency has been established through adopting the CEFR (Wu and Wu 2007). In Japan, a CEFR-J project was carried out over a period of 8 years, the result of which was a set of 647 descriptors for school learners of English in Japan (North 2014). According to North (2014) p. 59, one of the most interesting points about the CEFR-J is the breaking down of A1 into three levels, and there is a pre-A1 level. This shows that in adopting the existing CEFR, modifications or adaptations are necessary to suit the local context.

In addition to its influence on curricula/syllabi and assessment, the methods in developing descriptor pools, collecting data, and constructing and interpreting scales (North 2000) have proved of practical use and been adopted in similar projects.

To sum up, despite the fact that there have been criticisms on the CEFR over the lack of consideration of the construct underlying the original scales, and inconsistencies in terms of criteria and terms across levels (Vogt 2012, p. 49), the CEFR has achieved great success in that it has provided language educators and testers with a common framework for designing curricula, teaching materials, and tests. The experience gained from developing and implementing the CEFR in the European context has strong implications for language teaching/learning and assessment in other parts of the world.

TEM is a large-scale, nationwide test battery developed and administered by the National Advisory Committee for Foreign Language Teaching (NACFLT) on behalf of the Higher Education Department, Ministry of Education, P. R. China (Jin and Fan 2011). TEM consists of two levels, TEM4 and TEM8, and has now approximately an annual test population of 260,000 (TEM4) and 200,000 (TEM8) (TEM Test Office: Test report to universities, unpublished).

As is stated in the TEM test syllabus (2004), TEM has been designed to assess English language proficiency of undergraduate English majors to determine whether test takers have met the learning requirements specified in the teaching syllabus. Therefore, TEM may provide teachers with feedback on their teaching effectiveness,

and students on their strengths and weaknesses in English learning. In this sense, TEM is expected to facilitate the implementation of the teaching syllabus and to improve the quality of language teaching and learning for English majors across China NACFLT (2004a) and NACFLT (2004b).

One of the TEM components is writing, which amounts to 20% of the total score (TEM test syllabus 2004). The current TEM test report consists only of a total score, and no sub-score is provided for each part. Nor have descriptors for profiling TEM candidates' writing proficiency been included in the test report. As TEM is expected to provide feedback for teaching and learning, it has been deemed of great necessity for the test report to be more informative. To this end, writing was chosen as the research focus, and a study was launched, with the support from the TEM Test Office and the Foreign Language Teaching and Research Press, to develop a writing ability scale for TEM in addition to its existing assessment scale. The study aimed at providing a detailed profile of the test taker's writing proficiency for future use. During the process of the scale construction, various sources were made use of, including mainly, the TEM test syllabus, the teaching syllabus, and relevant CEFR descriptors. The inclusion of some CEFR descriptors was meant to enrich the descriptor pool, as the descriptors from the TEM test syllabus and the teaching syllabus were mainly academic study-oriented.

In the development of language proficiency scales, there are generally two categories of methods in use: intuitive methods and empirical methods (Fulcher 2003, p. 92). The former mainly includes expert judgments, committee, and experiential methods (Knoch 2009, pp. 43–44). In the latter group, descriptors are created through an empirically verifiable procedure and are based on observable learner behavior, including 'the data-based or data-driven approach,' 'empirically derived, binary-choice, boundary definition scales,' 'scaling descriptors' (Fulcher 2003, p. 92). The present study has used both intuitive and empirical methods in order to gain a better description of the subjects' writing ability.

This article describes and discusses the construction of a TEM writing ability scale, and attempts to address two research questions.

Research question one: compared with descriptors from other sources, to what extent are CEFR descriptors adaptable in describing TEM candidates' writing proficiency?

Research question two: how can we construct a writing ability scale of "can do" statements for TEM test takers by utilizing descriptors from various sources?

Methods

In order to answer the research questions, we adopted a mixed-method approach and designed a four-stage research procedure. In stage 1, we drafted a questionnaire consisting of "can do" statements. In stage 2, we conducted the questionnaire survey and built an initial descriptor pool based on the survey results. In stage 3, we invited teachers to judge the level of difficulty of the descriptors with reference to student writing samples with predetermined levels of proficiency. And in stage 4, we conducted small-scale interviews with a view to finalizing the writing ability scale profiling what TEM candidates can do at different levels of proficiency.

As has been mentioned above, two research instruments were used in the study: a "can do" questionnaire and a set of TEM writing scripts with predetermined logit values.

The main part of the “can do” questionnaire was composed of descriptors, the bulk of which came from B2 to C1 in the CEFR with a few descriptors from A1 and A2, and the current English teaching syllabus (2000) and the TEM test syllabus (2004), and focus group discussion. The decision to choose CEFR descriptors from levels B2–C1 was based on a comparison between the writing requirements in the current English teaching syllabus (2000), the TEM test syllabus (2004) and the relevant CEFR descriptors. And we found there was some comparability between the two in terms of the proficiency required for that level. For example, level 4 of the English teaching syllabus requires that students write a composition of 150–200 words within 30 min using topics, outlines or pictures, or graphs as prompts. The writing should be relevant to the topic, clear in structure and logic, correct in grammar, fluent, and appropriate in expression (NACFLT 2000: 10). In the CEFR’s “Reports and Essays,” level B2 is described as “Can write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem” (Council of Europe 2001, p. 62). We also included some A1 and A2 descriptors, mainly referring to writing letters and notes, in order to match those from the current English teaching syllabus (2000).

During the process of drafting the questionnaire, teachers were invited for a focus group discussion to comment on the typicality of the descriptors collected from the CEFR, the teaching or the test syllabus. Their comments would be considered when making decisions on the inclusion or exclusion of descriptors in the questionnaire. Besides, they were asked to write descriptors of typical writing activities which they considered important.

After several rounds of discussion and a pilot study, a questionnaire with 40 “can do” descriptors was finalized (see Appendix 1). The questionnaire used the 5-point Likert scale ranging from “not important at all” to “very important” for each descriptor, eliciting teacher participants’ opinions on its importance with reference to their teaching experience.

The second research instrument was a set of 36 student writing samples. They came from four operational administrations between 2011 and 2014, representing different levels of writing proficiency (see logit values in Appendix 2). The topics cover a range of domains, such as “Should Private Car Owners be Taxed for Pollution?,” “The Dragon Boat Festival,” “Coupon Collecting and Group-Buy Deals,” and “Should English Majors Study Maths?”

The subjects who participated in stages 2 and 3 were mostly teachers of writing, with 30 PhD students. One hundred ninety-four of them took part in stage 2 and 35 in stage 3.

In stage 2, 212 questionnaires were distributed to those teachers of writing and 30-odd PhD students in mid-June, 2014, and 194 were collected and proved to be valid.

To gain further evidence of the suitability of the CEFR in the local context, it was deemed necessary to investigate what level or levels these and other descriptors would fall into, which is the objective of stages 3 and 4. Stage 3 was the first part of the investigation, which consisted of teacher judgment made on the difficulty of a descriptor with relation to one particular student script. Thirty-five participants took part in

stage 3. During this stage, 36 TEM student scripts were first prepared, representing a range of proficiency levels (see Appendix 2), and then the 35 teachers were asked to read the scripts and complete the “can do” questionnaire (one for each script; see Appendix 3); that is, they were required to judge to what extent a student writer was competent with reference to the descriptors in the questionnaire. Thirty-five sets of “can do” questionnaires were distributed and returned valid. Then, we ran FACETS on the data in order to obtain the logit value of each descriptor. Here, the logit value can be understood as the difficulty (level) of a descriptor, i.e., a larger logit value corresponds to a higher level of ability. Appendix 1 presents the logit values obtained for the descriptors.

At stage 4, small-scale interviews were conducted to triangulate the obtained questionnaire data. To this end, eight university teachers were invited, between November 2014 and May 2015, to elicit their opinions on the difficulty level of the descriptors. The participants included three male teachers and five female teachers, seven of them being associate professors with experience of teaching writing ranging from 1 year to more than 10 years.

Results and discussion

Through several rounds of investigation, we obtained the data required for the construction of a writing ability scale for English majors. A reliability test was run on the 5-point Likert scale items in the questionnaire. The Cronbach's Alpha coefficients were 0.989, which indicates the reliability (inner consistency) was very satisfactory.

As can be seen from Appendix 1, the measures (difficulty value) of the descriptors range from -1.55 logits to $+1.66$ logits; the model standard error is 0.04; the Infit value ranges from 0.89 to 1.34, all falling in the range of 0.6 to 1.5, indicating that all the descriptors fit the model well. The average mean of the 40 descriptors in the questionnaire survey was 4.04 and the standard deviation stood at 0.26.

Among the 40 descriptors (Appendix 1), there were 16 with the mean value lower than 3.96 ($4.04 - 1.96 * 0.04$). We went through them one by one, and special attention was given to descriptors with the mean value lower than 3.60, i.e., nos. 23 and 37. Examining these two descriptors with relation to our classroom teaching practice, we found that English majors seldom write diaries nowadays and nor do they have the opportunity to write introductions to scenic spots. So, we decided to discard them in order to maintain a higher level of importance among the rest of the descriptors (see Appendix 1 for the remaining 38 descriptors). It can be noted that of the remaining 38 descriptors, 19 originate from the CEFR, ranging from A1 to C1, with B1 and B2 forming the majority. Of the other 19 descriptors, 13 come from the teaching or the test syllabus, and 6 from teacher contribution during the focus group discussion. The fact that all the CEFR inclusions have survived seems to suggest that the teachers regarded the descriptors as important indicators of writing ability. These 38 descriptors were then used to form a “can do” questionnaire to be used in the next stage (see Appendix 3).

It can be seen from Appendix 1 that five descriptors have a logit value of $+1$ (33, 34, 31, 32, and 36). Sixteen descriptors have logit values between 0 and 1 (30, 29, 17, 27, 26, 28, 35, 4, 15, 1, 3, 23, 14, 25, 5, and 2). And there are 17 descriptors with logit values below 0: 13, 18, 19, 38, 6, 20, 9, 24, 12, 10, 7, 8, 16, 22, 37, 21, and 11.

A close observation reveals some interesting points about the grouping. Descriptors with a logit value of +1 mostly refer to high-level academic writing activities. For example, can do some academic writing (33), can write a graduation thesis (31), can write research proposals (32), and can write reports of different genres (36). Descriptors with logits between 0 and 1 mostly point to routine teaching and learning activities: can write term papers (30), can write summaries based on argumentative texts (26), and can paraphrase the content of passages (23). Descriptors with logits below 0 are mostly simple writing activities, such as can write resumes (38), can write personal letters giving news (12), can write invitation letters appropriately (16), and can write self-introduction (37).

If we look at the origins of the descriptors, we can find that the 19 CEFR-related descriptors mainly fall below +1, with 7 between 0 and 1 and 12 below 0. Of the 13 teaching/test syllabus-related descriptors, 2 are above 1, 7 are between 0 and 1, and 4 are below 0. Of the six descriptors that came from the focus group discussion, three are above +1, 1 between 0 and 1, and 2 are below 0.

The following section analyzes and discusses the data with reference to our research questions.

Our first research question is the following: compared with descriptors from other sources, to what extent are CEFR descriptors adaptable in describing TEM candidates' writing proficiency? As shown by the statistics, descriptors of a relatively high difficulty (i.e., with logit value above 1) belong to the academic writing domain. For instance, can write abstract, literature review, etc. (33), can write a graduation thesis of 3000–5000 words (31), and can write research proposals (32). TEM test takers are supposed to be able to complete such tasks in their senior years. Those of a medium difficulty (i.e., logit value between 0 and 1) have a high level of similarity with classroom learning activities that teachers employ to improve students' writing proficiency. Some of them are as follows: can carry out the continuation task of writing after reading (27), can paraphrase the content of passages (23), and can write book reviews as required (29). By contrast, descriptors of a low difficulty (i.e., logit value below 0) refer to everyday simple writing activities and some of them belong to the personal/social domain. For example, can write short, simple formulaic notes relating to matters in areas of immediate need (7), can write personal letters/compositions describing experiences, feelings and events (9). Freshmen are expected to fulfill writing tasks of such nature NACFLT (2000). Thus, the three levels of difficulty, as defined by logit values, match the writing requirements stipulated in the teaching syllabus (2000) and teachers' experience, and can reflect the current teaching and learning status of writing proficiency.

As is mentioned before, of the 38 descriptors used in stage 3, 19 came from the CEFR covering five levels from A1 to C1. Of the five descriptors with logits above 1, none came from the CEFR, whereas among descriptors with logits between 0 and 1, 50% came from the CEFR, and among below 0 descriptors, 65% had a CEFR origin. Three points are worth mentioning here. The first point is the fact that the CEFR-related descriptors all survived the stage 2 questionnaire survey clearly indicates a high level of acceptance among the teachers in describing TEM candidates' writing proficiency.

Secondly, despite the apparent suitability of the CEFR-related descriptors, some of the levels previously assigned in the CEFR have changed in our study. And the biggest changes have occurred among descriptors between logits 0 and 1. Here are some examples. Descriptor no. 17 is assigned B1 in the CEFR, while descriptor no. 15 is originally at CEFR B2. In our study, the two descriptors have reversed their places, with no. 17 obtaining a higher logit value than no. 15 (see Appendix 1). Similar cases can also be found in other descriptors as well: the original C1 descriptor (no. 1) was found easier in our study than no. 15, which has an assigned CEFR level of B2.

Thirdly, when we observe the changes from the perspective of assigned CEFR levels, the biggest shifts have taken place between B1 and B2 descriptors; quite a few descriptors have changed positions.

There could be two reasons for the changes. First, the changes could be attributed to problems inherent in the original descriptors. Our data show that descriptor no. 4 (CEFR B1) has obtained a slightly higher logit value than descriptor no. 15 (CEFR B2). A close examination of the content of the two descriptors reveals that no. 4 actually describes a higher level of performance than no. 15. Common sense or teaching experience would tell us that “can write letters/compositions commenting on the correspondent’s news and views (No. 15)” would be relatively easier than “can convey information and ideas on abstract as well as concrete topics (No. 4).” As the original levels of descriptors seem to be a bit problematic, it is no wonder that they have swapped places in our data.

Secondly, the changes could be related to the teachers’ own perception of difficulty of writing activities which are embodied in the descriptors. As a result of their teaching/professional background, some teachers would consider certain writing activities more difficult than others. For instance, descriptor no. 27 (can carry out the continuation task of writing after reading) is considered a bit difficult with a logit value of 0.6. Even descriptor no. 28 (can describe a graph, chart, or table in detail) is viewed as possessing some difficulty (0.43). The difficulty value of the two tasks could be due to the fact that they are less frequently used and practiced in our writing classes, so that teachers are less familiar with the format and deem them difficult.

Although CEFR-related descriptors in our study have undergone significant shifts in their assigned levels, they have, generally speaking, proved to be feasible in profiling TEM candidates’ writing proficiency. For example, those descriptors in our scale with logits between 0 and 1 are mostly CEFR descriptors from B1 to C1; none of the level A descriptors have appeared in that range. Similarly, those CEFR-related descriptors with logit values below 0 on our constructed scale mainly come between levels A2 and B1; no CEFR-related descriptors with B2 and above have been included. That is, higher-level CEFR-related descriptors are located in a higher range on our scale; the same also applies to low-level CEFR-related descriptors which are grouped in the range below 0 on our scale. This answers our first research question: that is, some of the CEFR-related descriptors are, to some extent, adaptable in describing TEM test takers’ writing proficiency.

Our second research question is how can we construct a writing ability scale of “can do” statements for TEM test takers by utilizing descriptors from various sources? We employed library and survey research in building a “can do”

descriptor pool. First, we went through relevant literature and collected teaching and testing syllabuses and existing rating/proficiency scales. On the basis of an extensive review, we set up an initial descriptor pool by extracting descriptors from existing scales or by modifying or revising teaching requirements or testing objectives from the abovementioned teaching and testing syllabuses. A questionnaire was thus constructed from this descriptor pool. Then, we used survey methods to determine the relevance of the initial descriptors (stage 2) and the levels of the descriptors. Our experience has confirmed the advantage of using combined methods, as these methods have enabled us to improve the quality of the descriptors.

When all the descriptors were calibrated to estimated difficulties on a common logit scale as we had done in stage 3, the next step was to decide on the number of levels, or rather how to set the cut-off points on the scale. In this regard, North (2000), p. 293 has proposed three options. Option one is to “create a scale of more or less equal intervals.” Option two is to look for “patterns and clusters,” or “natural gaps” on the vertical scale. Option three is to match those “patterns and clusters” to generally accepted levels. Upon comparison, we found the third option, which combines with the second option, was more feasible in our context, as it could ensure objectivity and reliability in level identification while allowing us to check the validity of the identified levels against the existing teaching and test syllabuses.

By means of the third method, we carefully reviewed the vertical scale consisting of the 38 descriptors (see Appendix 1), and identified two natural gaps. The first gap appeared between nos. 28 and 35, as there was a difference of 0.17 logit between the two, which was larger than the differences between the adjacent descriptors. The second gap was located between nos. 13 and 18, with a logit difference of 0.18, again bigger than the difference between any other nearby descriptors. Upon review and discussion, it was decided to use the two natural gaps as the cut-off points on the scale. Thus, the 38 descriptors were divided into three levels (see Appendix 4).

A draft version of the scale was thus obtained. However, the three levels were determined mainly statistically at this stage. In order to testify the validity of our categorization, more evidence was needed to triangulate the existing data. To this end, eight university teachers were invited to read the draft version of the scale carefully and then comment on the descriptors based on his/her intuition, teaching experience, and rating experience if any. If they disagreed with the level categorization of a given descriptor, they were asked to provide explanation and also suggest a new level for that particular descriptor.

Based on the content analysis of the interview data, some modifications were made to the draft scale ranging from modification of wording to level adjustment. For example, “can write speech scripts for *presentations*” was specifically defined as “can write speech scripts for *classroom presentations*”; “can express oneself with clarity and precision in writing” was simplified as “can clearly and precisely express oneself”; the two original level 1 descriptors “can take messages communicating enquiries” and “can take messages explaining problems” were combined into one descriptor “can take messages communicating enquiries, explaining problems” and was upgraded to level 2, and “can describe a graph, chart or table in detail,” which was originally placed at level 3 was moved down to level 2.

Thus, we finalized the writing ability scale for English majors in Chinese universities as seen in Table 1.

In the final version of the writing ability scale, level 1 contains 14 descriptors; level 2, 13 descriptors; and level 3, 10 descriptors (totaling 37 descriptors with two of the

Table 1 Writing ability scale for English majors in Chinese universities in the final version of the writing

Level 3: 10 descriptors

- can do academic writing (e.g. abstract, introduction, literature review, methodology, data findings and analysis, conclusion)
- can proofread and give feedback to the academic writing of students of other majors
- can write a graduation thesis
- can write research proposals applying for further education abroad
- can write reports of different genres (e.g. business report)
- can write term papers (research papers written by students over an academic term)
- can write book reviews as required
- can express thoughts about cultural topics such as music, films in writing
- can carry out the continuation task of writing after reading the given part
- can write summaries based on augmentative texts

Level 2: 13 descriptors

- can describe a graph, chart or table in detail
- can write speech scripts for classroom presentations
- can convey information and ideas on abstract as well as concrete topics
- can write letters/compositions commenting on the correspondent's views
- can clearly and precisely express oneself
- can effectively express one's ideas, emotions
- can paraphrase the content of the English passages
- can write letters/compositions highlighting the personal significance of events and experiences
- can write summaries based on narrative and descriptive texts
- can check information (e.g. examining general patterns of information across text types)
- can express news and views effectively in writing
- can write letters/compositions conveying degrees of emotion
- can take messages communicating inquiries, explaining problems

Level 1: 14 descriptors

- can write resumes
- can make and respond to inquiries in writing
- can provide information needed for registrations of all kinds (e.g. hotel and website registration)
- can write personal letters/compositions describing experiences, feelings and events
- can write a short passage (150–200 words) based on a given topic or outline
- can write personal letters/compositions giving news
- can ask for or pass on personal details in written form
- can write short, simple formulaic notes relating to matters in areas of immediate need
- can write personal letters and notes conveying information
- can write invitation letters
- can write notices
- can write self-introductions
- can write to communicate with overseas friends through letters or e-mails
- can write very simple personal letters expressing thanks and apology

original 38 descriptors combined). Comparing the divisions on the scale with the English teaching syllabus (2000), we have found that level 3 descriptors generally match the ability level required for senior students of English, level 2 descriptors correspond to requirements for junior students of English, while those of level 1 are mostly what freshmen majoring in English are supposed to be competent in. This means that the scale can cover a relatively wide range of writing proficiency of those students who take TEM.

To sum up, the present study serves two purposes: (1) to investigate the suitability of CEFR descriptors, among other descriptor sources, in profiling the writing proficiency of TEM test takers, who are university students, and (2) to explore how to construct a writing ability scale in the local context. Our findings show that some CEFR-related descriptors are generally suitable for describing the writing proficiency of the target test population, though their new levels on the constructed scale have differed, in many cases, from their original levels. In scaling the descriptors obtained from the third stage of the study, we adopted one of the methods proposed by North (2000), p. 293 by looking for “natural gaps” or “clusters of data” in the FACETS measurement scale, and then we compared the gaps with the requirements stipulated in the teaching syllabus (2000). On the basis of the comparison and the results of the following interviews, we decided on two cut-off points that divided the descriptors into three levels. The scaled descriptors generally correspond to what the teaching syllabus has stipulated.

Conclusions

The present study investigates into the development of a writing ability scale in the context of TEM. The immediate purpose of the study is to further improve the validity of the TEM writing sub-test by providing test feedback which is more informative in that it can offer a writing ability profile for test candidates. It is our strong belief that the constructed writing ability scale, despite the fact that it needs further improvement, is expected to facilitate classroom teaching of writing, thus achieving our aim of assessment for learning.

Besides its immediate practical value for TEM writing tests, the present study is expected to have implications in two aspects. One is that it provides evidence for the adaptability of CEFR-related descriptors in a Chinese higher education context, which has seldom been researched into so far, and thus it can shed light on further study in other ability domains such as reading, listening, speaking, and translation, which could result in a more comprehensive language ability profile of English language majors. Those ability scales are supposed to help teachers in teaching and evaluation, and could be instrumental in test design and the construction of marking scales.

The other implication lies in the fact that the study provides exemplary practice for developing scales of a similar kind. Our study has adopted a research approach that combines qualitative methods with quantitative ones. And the final product has resulted from survey study, teacher judgment, and statistical analysis. Use of data from multiple sources has proved to be instrumental in our study, which may be useful to researchers in similar studies in the future.

Appendix 1

Table 2 Mean value, source, logit value, S.E., and Infit MnSq. of the descriptors (pilot version and revised version)

| No. (pilot version) | Descriptors | Mean value | No. (revised version) | Source | Logit value | S.E. | Infit MnSq. |
|---------------------|--|------------|-----------------------|------------------------|-------------|------|-------------|
| 1 | can express oneself with clarity and precision in writing | 4.68 | 1 | CEFR-C1 | 0.18 | 0.04 | 0.95 |
| 2 | can express news and views effectively in writing | 4.35 | 2 | CEFR-B2 | 0.01 | 0.04 | 0.90 |
| 3 | can express one's ideas, emotions effectively in writing | 4.37 | 3 | CEFR-B2 | 0.15 | 0.04 | 0.93 |
| 4 | can convey information and ideas on abstract as well as concrete topics | 4.17 | 4 | CEFR-B1 | 0.24 | 0.04 | 0.96 |
| 5 | can check information through writing (e.g. examining general patterns of information across text types) | 3.93 | 5 | CEFR-B1 | 0.05 | 0.04 | 0.97 |
| 6 | can ask about or explain problems by writing | 4.16 | 6 | CEFR-B1 | -0.46 | 0.04 | 0.93 |
| 7 | can write short, simple formulaic notes relating to matters in areas of immediate need | 4.07 | 7 | CEFR-A2 | -0.77 | 0.04 | 1.04 |
| 8 | can write personal letters and notes asking for or conveying information | 4.13 | 8 | CEFR-B1 | -0.92 | 0.04 | 1.00 |
| 9 | can write personal letters/compositions describing experiences, feelings and events | 4.06 | 9 | CEFR-B1 | -0.58 | 0.04 | 0.97 |
| 10 | can ask for or pass on personal details in written form | 3.83 | 10 | CEFR-A1 | -0.76 | 0.04 | 1.06 |
| 11 | can write very simple personal letters expressing thanks and apology | 4.00 | 11 | CEFR-A2 | -1.55 | 0.04 | 1.13 |
| 12 | can write personal letters giving news | 3.68 | 12 | CEFR-B1 | -0.67 | 0.04 | 0.95 |
| 13 | can write letters/compositions conveying degrees of emotion | 3.70 | 13 | CEFR-B2 | -0.01 | 0.04 | 0.93 |
| 14 | can write letters /compositions highlighting the personal significance of events and experiences | 3.80 | 14 | CEFR-B2 | 0.12 | 0.04 | 0.99 |
| 15 | can write letters/compositions commenting on the correspondent's news and views | 3.79 | 15 | CEFR-B2 | 0.2 | 0.04 | 0.94 |
| 16 | can write invitation letters appropriately | 4.00 | 16 | CEFR-A2 | -1.02 | 0.04 | 1.09 |
| 17 | can express thoughts about cultural topics such as music, films in writing | 3.81 | 17 | CEFR-B1 | 0.75 | 0.04 | 1.05 |
| 18 | can take messages communicating enquiries | 3.91 | 18 | CEFR-B1 | -0.19 | 0.04 | 0.89 |
| 19 | can take messages explaining problems | 3.99 | 19 | CEFR-B1 | -0.36 | 0.04 | 0.92 |
| 20 | can provide information needed for registrations of all kinds (e.g. hotel and website registration) | 3.96 | 20 | Focus group discussion | -0.49 | 0.04 | 1.28 |
| 21 | can write to communicate with overseas friends through letters or E-mails | 4.07 | 21 | Focus group discussion | -1.27 | 0.04 | 1.02 |
| 22 | can write notices | 3.95 | 22 | Teaching syllabus | -1.18 | 0.04 | 1.06 |
| 23 | can write diaries | 3.53 | deleted | Focus group discussion | X | X | X |
| 24 | can paraphrase the content of passages | 3.92 | 23 | Teaching syllabus | 0.14 | 0.04 | 0.93 |
| 25 | can write a short passage based on a given topic outline | 4.25 | 24 | Teaching syllabus | -0.62 | 0.04 | 0.90 |
| 26 | can write summaries based on narrative and descriptive texts | 4.27 | 25 | Teaching syllabus | 0.05 | 0.04 | 0.94 |

Table 2 Mean value, source, logit value, S.E., and Infit MnSq. of the descriptors (pilot version and revised version) (*Continued*)

| No. (pilot version) | Descriptors | Mean value | No. (revised version) | Source | Logit value | S.E. | Infit MnSq. |
|---------------------|--|------------|-----------------------|------------------------|-------------|------|-------------|
| 27 | can write summaries based on augmentative texts | 4.27 | 26 | Teaching syllabus | 0.56 | 0.04 | 0.92 |
| 28 | can carry out the continuation task of writing after reading | 3.77 | 27 | Teaching syllabus | 0.6 | 0.04 | 0.90 |
| 29 | can describe a graph, chart or table in detail | 3.99 | 28 | Test syllabus | 0.43 | 0.04 | 0.98 |
| 30 | can write book reviews as required | 3.95 | 29 | Teaching syllabus | 0.85 | 0.04 | 0.95 |
| 31 | can write term papers | 4.26 | 30 | Teaching syllabus | 0.91 | 0.04 | 0.90 |
| 32 | can write a graduation thesis | 4.45 | 31 | Teaching syllabus | 1.4 | 0.04 | 0.94 |
| 33 | can write research proposals applying for further education abroad | 4.26 | 32 | Focus group discussion | 1.33 | 0.04 | 1.05 |
| 34 | can do some academic writing (e.g. abstract, introduction, literature review, methodology, data findings and analysis, conclusion) | 4.13 | 33 | Focus group discussion | 1.66 | 0.04 | 1.17 |
| 35 | can proofread and give feedback to the academic writing of students of other majors | 3.75 | 34 | Focus group discussion | 1.46 | 0.04 | 1.34 |
| 36 | can write speech scripts for presentations | 4.08 | 35 | Focus group discussion | 0.26 | 0.04 | 0.97 |
| 37 | can write appropriate introductions to the scenic spots | 3.59 | deleted | Focus group discussion | X | X | X |
| 38 | can write reports of different genres (e.g. business report) | 3.81 | 36 | Teaching syllabus | 1.08 | 0.04 | 0.98 |
| 39 | can write self-introductions | 4.42 | 37 | Teaching syllabus | -1.2 | 0.04 | 1.03 |
| 40 | can write resumes | 4.50 | 38 | Teaching syllabus | -0.38 | 0.04 | 1.23 |

Appendix 2

Table 3 Logit values of TEM4 writing scripts (2011–2014)

| Order | 2011 | | 2012 | | 2013 | | 2014 | |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|
| | St. no. | Measure | St. No. | Measure | St. no. | Measure | St. no. | Measure |
| 1 | 18* | 2.71 | 14* | 1.93 | 10* | 1.73 | 14* | 1.55 |
| 2 | 25* | 2.54 | 16* | 1.73 | 1* | 1.50 | 16* | 1.46 |
| 3 | 24* | 2.28 | 3* | 1.53 | 7* | 1.30 | 3* | 1.21 |
| 4 | 26 | 2.08 | 15 | 1.47 | 16 | 1.19 | 2 | 1.13 |
| 5 | 14 | 1.84 | 10 | 1.33 | 4* | 1.09 | 4 | .94 |
| 6 | 15 | 1.66 | 5* | 1.32 | 5* | 1.04 | 17* | .89 |
| 7 | 17 | 1.61 | 4* | 1.20 | 14* | .72 | 19* | .45 |
| 8 | 28* | 1.60 | 17* | 1.19 | 2 | .69 | 11* | .41 |
| 9 | 16* | 1.55 | 8 | .76 | 11* | .67 | 1 | .39 |
| 10 | 8* | 1.51 | 2* | .72 | 8* | .39 | 5* | .36 |
| 11 | 22 | 1.42 | 18* | .53 | 9 | .24 | 18* | .27 |
| 12 | 12 | 1.10 | 20 | .30 | 6 | .15 | 20 | .16 |
| 13 | 21 | 1.08 | 12 | .19 | 12 | .08 | 13* | -.01 |
| 14 | 1* | .87 | 11 | .15 | 13* | -.02 | 10 | -.03 |
| 15 | 3* | .80 | 13 | .12 | 3 | -.29 | 6 | -.05 |
| 16 | 23 | .65 | 19 | .03 | 15 | -.73 | 15 | -.08 |
| 17 | 11 | .63 | 6* | -.14 | | | 8 | -.20 |
| 18 | 6 | .44 | 1 | -.48 | | | 9 | -.41 |
| 19 | 4 | .30 | 9 | -.72 | | | 12 | -.47 |
| 20 | 27 | .27 | 7 | -1.82 | | | 7 | -1.01 |
| 21 | 5 | .24 | | | | | | |
| 22 | 19 | .14 | | | | | | |
| 23 | 13* | -.05 | | | | | | |
| 24 | 7 | -.08 | | | | | | |
| 25 | 10 | -.33 | | | | | | |
| 26 | 20 | -.70 | | | | | | |
| 27 | 9 | -.94 | | | | | | |
| 28 | 2 | -1.52 | | | | | | |

The scripts written by students with "*" are selected and used in stage 3 of the study. The criteria of essay selection were adopted as follows:

1) 9 essays of each year covering four levels were selected;

2) Level distribution was three essays for level 1, another three for level 2, two for level 3, and one for level 4

3) The 4 levels were identified as follows: those with the three highest logit values in each administration were classified into level 1. The two with the logit values closest to the mean (one was higher and the other lower) were classified into level 3. The three essays in the very middle of level 1 and level 3 were classified into level 2. If the number of the essays between levels 1 and 3 was odd, then the one with a smaller logit value was chosen, as in the case of 2011 and 2012.

The one with a logit value slightly lower than 0 was classified into level 4

Appendix 3

Table 4 A “can do” questionnaire (stage 3)

| Descriptors | 1 | 2 | 3 | 4 | 5 |
|--|---|---|---|---|---|
| 1. can express oneself with clarity and precision in writing | | | | | |
| 2. can express news and views effectively in writing | | | | | |
| 3. can express one's ideas, emotions effectively in writing | | | | | |
| 4. can convey information and ideas on abstract as well as concrete topics | | | | | |
| 5. can check information through writing (e.g. examining general patterns of information across text types) | | | | | |
| 6. can ask about or explain problems by writing | | | | | |
| 7. can write short, simple formulaic notes relating to matters in areas of immediate need | | | | | |
| 8. can write personal letters and notes asking for or conveying information | | | | | |
| 9. can write personal letters/compositions describing experiences, feelings and events | | | | | |
| 10. can ask for or pass on personal details in written form | | | | | |
| 11. can write very simple personal letters expressing thanks and apology | | | | | |
| 12. can write personal letters/compositions giving news | | | | | |
| 13. can write letters/compositions conveying degrees of emotion | | | | | |
| 14. can write letters/compositions highlighting the personal significance of events and experiences | | | | | |
| 15. can write letters/compositions commenting on the correspondent's news and views | | | | | |
| 16. can write invitation letters appropriately | | | | | |
| 17. can express thoughts about cultural topics such as music, films in writing | | | | | |
| 18. can take messages communicating enquiries | | | | | |
| 19. can take messages explaining problems | | | | | |
| 20. can provide information needed for registrations of all kinds (e.g. hotel and website registration) | | | | | |
| 21. can write to communicate with overseas friends through letters or E-mails | | | | | |
| 22. can write notices | | | | | |
| 23. can paraphrase the content of passages | | | | | |
| 24. can write a short passage based on a given topic or outline | | | | | |
| 25. can write summaries based on narrative and descriptive texts | | | | | |
| 26. can write summaries based on augmentative texts | | | | | |
| 27. can carry out the continuation task of writing after reading | | | | | |
| 28. can describe a graph, chart or table in detail | | | | | |
| 29. can write book reviews as required | | | | | |
| 30. can write term papers | | | | | |
| 31. can write a graduation thesis | | | | | |
| 32. can write research proposals applying for further education abroad | | | | | |
| 33. can do some academic writing (e.g. abstract, introduction, literature review, methodology, data findings and analysis, conclusion) | | | | | |
| 34. can proofread and give a feedback to the academic writing of students of other majors | | | | | |
| 35. can write speech scripts for presentations | | | | | |
| 36. can write reports of different genres (e.g. business report) | | | | | |
| 37. can write self-introductions | | | | | |
| 38. can write resumes | | | | | |

(Please read the script and then judge to what extent the student writer is competent in performing what the descriptor says: 1 = not competent at all; 2 = less competent; 3 = competent; 4 = more competent; 5 = very competent)
Thanks for your contribution!

Appendix 4

Initial calibration of the scale

Level 3

- 33. can do some academic writing (e.g. abstract, introduction, literature review, methodology, data findings and analysis, conclusion)
- 34. can proofread and give a feedback to the academic writing of students of other majors
- 31. can write a graduation thesis
- 32. can write research proposals applying for further education abroad
- 36. can write reports of different genres (e.g., business report)
- 30. can write term papers
- 29. can write book reviews as required
- 17. can express thoughts about cultural topics such as music, films in writing
- 27. can carry out the continuation task of writing after reading
- 26. can write summaries based on augmentative texts
- 28. can describe a graph, chart or table in detail

Level 2

- 35. can write speech scripts for presentations
- 4. can convey information and ideas on abstract as well as concrete topics
- 15. can write letters/compositions commenting on the correspondent's news and views
- 1. can express oneself with clarity and precision in writing
- 3. can express one's ideas, emotions effectively in writing
- 23. can paraphrase the content of passages
- 14. can write letters/compositions highlighting the personal significance of events and experiences
- 25. can write summaries based on narrative and descriptive texts
- 5. can check information through writing (e.g. examining general patterns of information across text types)
- 2. can express news and views effectively in writing
- 13. can write letters/compositions conveying degrees of emotion

Level 1

- 18. can take messages communicating enquiries
- 19. can take messages explaining problems
- 38. can write resumes
- 6. can ask about or explain problems by writing
- 20. can provide information needed for registrations of all kinds (e.g., hotel and website registration)
- 9. can write personal letters/compositions describing experiences, feelings and events
- 24. can write a short passage based on a given topic or outline
- 12. can write personal letters/compositions giving news
- 10. can ask for or pass on personal details in written form
- 7. can write short, simple formulaic notes relating to matters in areas of immediate need
- 8. can write personal letters and notes asking for or conveying information
- 16. can write invitation letters appropriately
- 22. can write notices.

37. can write self-introductions.

21. can write to communicate with overseas friends through letters or E-mails

11. can write very simple personal letters expressing thanks and apology

Acknowledgements

The authors would like to thank FLTRP and MOE for the funding they provided for this research. The authors are also grateful to all the teachers who kindly took part in the survey and the rating. Finally, our appreciation goes to the anonymous reviewers of this article for the extremely insightful feedback they provided.

Funding

This paper is based on the project "The Applicability of the CEFR for English Language Education in China" (GZ20140100) funded by the Foreign Language Teaching and Research Press and the Key Project of Philosophy and Social Sciences "The Development of China Standards of English" (15JZD049) funded by the Ministry of Education, P. R. China.

Authors' contributions

Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Author details

¹School of English Studies, Shanghai International Studies University, No.550, Dalian Rd (W), Shanghai 200083, People's Republic of China. ²School of Foreign Languages, Jiangxi Science and Technology Normal University, No. 589 Xuefu Rd, Hongjiaozhou, Nanchang 330031, Jiangxi, People's Republic of China.

Received: 15 June 2017 Accepted: 13 October 2017

Published online: 15 November 2017

References

- Council of Europe (1996a). Common European framework of reference for language learning and teaching. In *Draft 1 of a framework proposal*. Strasbourg: Council of Europe.
- Council of Europe (1996b). Modern languages: Learning, teaching, assessment. A common European framework of reference. In *Draft 2 of a framework proposal*. Strasbourg: Council of Europe.
- Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Council of Europe (2003). *Relating language examinations to the common European framework of reference for languages: learning, teaching, assessment. Manual, preliminary pilot version*. Strasbourg: Council of Europe.
- Council of Europe (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.
- Council of Europe (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. Strasbourg: Council of Europe.
- Fulcher, G (2003). *Testing second language speaking*. London: Longman/Pearson Education.
- Huhta, A, Luoma, S, Oscarson, M, Sajavaara, K, Takala, S, Teasdale, A (2002). DIALANG—a diagnostic language assessment system for learners. In JC Alderson (Ed.), *Common European framework of reference for languages: Learning, teaching, assessment. Case studies*, (pp. 130–145). Council of Europe: Strasbourg.
- Jin, Y, & Fan, J. (2011). Test for English majors (TEM) in China. *Language Testing*, 28(4), 589–596.
- Khalifa, H. & French, A. (2008). Aligning Cambridge ESOL examinations to the CEFR: Issues and practice. Paper presented at the 34th Annual Conference, International Association for Educational Assessment.
- Knoch, U (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. New York: Peter Lang.
- Martyniuk, W. (2012). The use and (potential) misuse of frameworks—the CEFR case. European Centre for Modern Languages (ECML) of the Council of Europe, Austria: Graz. Paper presented at the ACTFL-CEFR Symposium 2012, Graz, Austria.
- NACFLT (2000). *Syllabus for university English language teaching*. Shanghai: Shanghai Foreign Language Education Press.
- NACFLT (2004a). *Syllabus for TEM4*. Shanghai: Shanghai Foreign Language Education Press.
- NACFLT (2004b). *Syllabus for TEM8*. Shanghai: Shanghai Foreign Language Education Press.
- North, B (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.
- North, B (2002). A CEF-based self-assessment tool for university entrance. In JC Alderson (Ed.), *Common European framework of reference for languages: Learning, teaching, assessment. Case studies*, (pp. 87–105). Council of Europe: Strasbourg.
- North, B (2014). *The CEFR in practice*. Cambridge: Cambridge University Press.
- Vogt, K (2012). Adaptations of CEFR descriptors in local contexts. In D Tsagarí, I Csépes (Eds.), *Collaboration in language testing and assessment*. New York: Peter Lang.
- Wu, J. R. W. & Wu, Y. F. (2007). Using the CEFR in Taiwan: the perspective of a local examination board. The Language Training and Testing Center, Taipei: Taiwan. Mimeo, paper presented at EALTA conference, Sitges, Spain.