Language Testing in Asia

RESEARCH Open Access

Looking beyond scores: validating a CEFR-based university speaking assessment in Mainland China

Li Liu* and Guodong Jia

* Correspondence: liliu@ruc.edu.cn
School of Foreign Languages,
Minde International Building,
Renmin University of China, No. 59
Zhongguancun Street, Haidian
District, Beijing, China

Abstract

Background: The present study examined the validity of a university-based speaking assessment (Test of Oral Proficiency in English, TOPE for short) in mainland China. The speaking assessment was developed to meet the standards (Standard for Oral Proficiency in English, SOPE for short) set for teaching and learning of the oral English by the university.

Methods: The degree of interaction among candidates in the second part of the test (Presentation and Discussion) was analyzed in terms of frequency of the language functions. These functions were reflected in the test syllabus, which was developed based on the CEFR.

Results: Quantitative analysis revealed that majority of language functions intended by the test syllabus has been elicited in candidate performances. At the same time, students showed a relatively lack of interactional functions. Further qualitative investigations identified some speaking features that have the potential to affect scores of the candidate.

Conclusions: Merits and limitations of using the CEFR for developing the speaking assessment (i.e., applicability of the CEFR for TOPE) were discussed. The study was concluded with its limitations and suggestions for future study.

Keywords: Speaking assessment, TOPE, Language function, Speaking feature

Background

Performance assessment nowadays has been an integral part in the evaluation of student language learning. At the same time, a test is both a learning experience and a reflection of student learning (Mendelsohn 1989). Localized language assessment is therefore necessary to cater for language needs and to examine language proficiency of different groups of learners. In mainland China, a large number of English language assessments have been developed by educational and examinations authorities at different levels to meet the demand of the huge and ever-growing English learner population (Cheng & Curtis 2010), for example, The Tsinghua English Proficiency Test (TEPT) (Tsinghua English Proficiency Test Group 2012) and The Fudan English Test (Fan & Ji 2013).

Recent years have also brought language testers in China a much clearer understanding of the nature of L2 language proficiency and of the many different research approaches. And speaking has always been one of the major concerns for researchers and teachers.

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Various strands of studies have thus been conducted to help us in investigating complex issues of speaking tests. Among these studies, the development of assessment criteria and rating scales (or band descriptors) is clearly an important focus (e.g, Jin et al 2008; Sun & Wei 2012; Yi & Zeng 2004). Collection and analysis of test takers' scores and discourse features are also one of the major focuses of the research to provide evidence to candidate as well as rater performance (e.g., Pan 2013; Wang 2015). Spoken output among candidates and between candidates and examiners is also the subject of ongoing research (e.g., Tang 2014). However, few endeavors have been devoted to developing and validating university-based speaking assessment.

Under this backdrop, Test of Oral Proficiency in English (TOPE) is developed and administered by Renmin university of China to assess the oral communicative proficiency of college students who have undertaken compulsory English courses. The design of the test is driven by the communication needs of the college students and communicative competence defined in domestic and internationally recognized guidelines, namely, the College English Curriculum Requirements (Ministry of Education 2007) and the Common European Framework of References (CEFR for short, Council of Europe 2001). The theoretical correspondence between the TOPE and the CEFR is realized in the way that the Common Reference Levels are referred to in order to develop the level descriptors of the rating scale;qualitative aspects of spoken language use are drawn upon to provide guidance to the content of the test, including task setting parameters, e.g., response formats (monologic and dialogic), channels of communication (aural, written, graphic, visual) and discourse modes (description, argumentation, persuasion). In this regard, students are expected to use English to deal with real-world issues. The speaking construct is defined in the degree to which students can express their ideas in both monologic and interactive ways and collaborate in the creation and sustaining of interaction. The TOPE therefore adopts both one-to-one and the paired candidate formats based not only upon pedagogical considerations, but also upon the findings of various studies of spoken language discourse. And the test comprises a multi-task design, which allows the elicitation of different patterns of spoken interaction.

Specifically, the TOPE is consisted of three sections. Section A is Reading aloud and Retelling. Section B is Presentation and Discussion. Section C is Impromptu Speech and Question and Answer (see Table 1 for details). These tasks are expected to cater for different levels of cognitive demands and to elicit the types of language functions that are considered important by test developers and language teachers at Renmin University of China. Accordingly, tasks are designed to elicit language functions that are congruent with the results of questionnaires of teachers and students as well as those that are considered relevant based on the literature review by the project team. Results of the questionnaire study showed the needs to improve oral English of students (84.96%) and understandings of oral proficiency by both teachers and students (TOPE Project Group 2013). And the test defines speaking proficiency into six levels with level six being the highest.

As reflected in the construct definition of the TOPE, speaking is regarded as both a cognitive and a social endeavor. The design of the test shows the understanding that language communication comprises reciprocity and depends on the negotiation of meaning and joint co-construction between two or more persons (Jacoby & Ochs 1995). The similar view is also manifested in the CEFR in the division between

Table 1 The Component and Structure of TOPE

Part	Task	Test format	Time	Weighting	Can-do requirements
1	Reading-aloud and retelling	One student at a time	5 m	20%	-Can read English essays/stories aloud correctly in appropriate pronunciation and intonation. -Can retell the audio/video materials in a coherent, complete, standard and appropriate way. -Can be confident in using retelling strategies
2	Presentation and discussion	Two or three students in a group	18 m	45%	-Can give brief information about own family, living conditions and interest, etc. - Can give clear description and presentation according to the given visual or written prompt (e.g., picture, charts or photos) in a thoughtful, fluent, logic and well-structured way. -Can converse comfortably and appropriately on relevant topics. Can convey explicit ideas and opinions and respond to others' argument convincingly. -Can communicative effectively by employing both verbal and non-verbal strategies.
3	Impromptu speech and Q&A	One student at a time	5 m	35%	-Can produce clear, smoothly flowing, and well-structured speech with an effective logical structure. -Can respond to questions relevant to the topic adequately and precisely. -Can use speech and Q&A strategies appropriately.

'production' and 'interaction' to speaking (Council of Europe 2001). In addition, the CEFR (2001) places primary focus on language as a means of communication, and consequently attach great importance to the role of language functions. Language learners are therefore grouped into different proficiency levels in terms of what they can do with language, rather than the ability to handle specific grammatical structures or lexical items. As the CEFR suggests, the TOPE also features a qualitative progression in the functions that characterize learner language among levels.

One point needing to be mentioned is that the CEFR places great emphasis on language as a means of communication, and consequently puts language functions in a central position. In line with the CEFR, the TOPE espouses the leading role of functions and the speaking tasks are thus designed to explicitly elicit the language functions which students can demonstrate at various proficiency levels. Adopting this perspective of language function, the study therefore aims to examine features of context validity of the TOPE, as outlined in Weir's (2005, p. 46) test validation framework. The framework comprises a number of components each of which must be attended to by the test developer at the validation stage. As one of the major components, context validity concerns the contextual parameters of a test, which are often socially or externally determined in terms of the tasks with its specified lexical, structural and functional input and expected output.

To put it specifically, the study examined how well the TOPE test elicits intended language functions that learners are required to demonstrate. It is expected that students will differ across proficiency levels in terms of the range of exponents which they can use to perform those functions and by the degree of accuracy and complexity with which they can express their view. Previous studies exploring language functions have pre-defined speaking features that distinguished proficiency levels and examined these features using a quantitative approach (Brown 2006; Young 1995). Few studies have looked closely into the data for differences using a qualitative method. The current study thus attempts to distinguish the differences in language functions and speaking features from a mixed-methods perspective.

Specifically, the current study investigated the following research questions:

1. To what extent does the TOPE elicit intended language functions?
2. Which speaking features distinguish proficiency at each level?

To answer the first research question, language function analysis is conducted. The second research question is addressed by looking into the speaking performance of students qualitatively using conversation analysis. Speaking feature is operationalized as features of candidate's oral production within a turn that has the potential to lead to a change in a candidate's score (Seedhouse et al. 2014)

To narrow down the scope of the study, the Section B of the test was chosen as a focus for investigation. This section takes the form of group interview and discussion with three or four candidates in a group. First, each candidate was required to give a short self-introduction within one minute. Second, the interlocutor gave a prompt card to each of the candidates. The prompts are sentences, pictures or tables/graphs, and all relate to the same topic for discussion. After a short preparation (1 min), each candidate in turn made an individual presentation (2 min per candidate) according to the prompt, which intends to examine whether the candidate has the ability to describe and present information effectively. Third, the interlocutor asked all three candidates to participate in an eight-minute group discussion to elaborate further on the given topic. Fourth, each candidate answered one question raised by the interlocutor. In this section of the test, candidates are expected to participate, express opinions and negotiate with one another in discussion with a view to reaching an agreement or arriving at a conclusion.

Methods

Participants

The study involved 54 university students, two trained interlocutors and two experienced raters. The students were recruited from different departments at Renmin University of China, so as to cover a wide range of proficiency levels since students are admitted by different departments in terms of their scores of College Entrance Exam (Gaokao). They were all first-year students at the time of study who had spent about eight months at the university at the time of data collection.

Two examiners were present during each test and they shouldered different responsibilities. The interlocutor conducted the test, interacted with the test takers, and provided a global assessment of each test taker's performance using a holistic rating scale. The second examiner acted as an observer, who took no leading part in the test, but focused on making an assessment of each test taker based on a set of analytic scales. Based on this scoring method, two interlocutors were recruited from English language teachers at the university. They all had attended an interlocutor training sessions prior to the test event. The two raters were experienced teachers with experiences as professional raters in standardized speaking test.

Data collection and analysis

All performances were video-recorded, and the raters using the rating scale described earlier rated all the recorded performances. The rating scale of the Section B covers four analytical dimensions, including Content, Delivery, Interactive communication and Global performance (Renmin University of China, 2013b). Original scores of the performances were not selected since unfit raters were identified based on the results of rater performance analysis conducted after the live test. And a total of eight discussion topics were used.

Transcribing the video-recorded performance

All video recordings were transcribed using a slightly simplified version of conversation analysis notation (Atkinson & Heritage 1984). Conversation analysis (CA) transcription was chosen in that it is informative and enables us to examine micro-features of interaction exhibited among the candidates.

Two research assistants, who had been trained to transcribe speaking data, transcribed the recordings. The researchers carefully checked selected portions of the data, and some modifications were suggested before the completion of the rest of transcriptions. An interactive, consensus approach was taken to ensure consistency in transcriptions. Prior to the segmentation of the transcripts, all transcripts were double-checked by the researchers while watching all speaking recorded samples.

Language function analysis

The language functions outlined in the CEFR and O'Sullivan et al. (2002) observation checklist were referred to for coding with the transcription (see Table 2). The useful tripartite distinction of functional resources between informational, interactional and interaction management functions were remained (O'Sullivan et al. 2002). The checklist was revised after a preliminary coding drawing upon the established checklist with a small proportion of data. The revision included the combination of some functions (e.g., Expressing opinions/preferences) and removal of certain function from the list, such as Reciprocating. The revised checklist thus consisted of a clear distinction of *informational* (e.g., expressing opinion, justifying opinion), *interactional* (e.g., asking for information, negotiating meaning), and *managing interaction* functions (e.g., initiating, reciprocating). The checklist was used in that it was originally developed for analyzing language functions elicited from candidates in paired speaking tasks of the Cambridge Main Suite examinations. And successful applications of the list to other speaking tests has already been empirically explored (Brooks 2003).

The selection of the language functions was also drawn partly on the requirements of the College English Curriculum Requirements (Ministry of Education 2007) which serves as the guidance for conducting college English teaching and learning in mainland China. The functions covered by the Requirements include Disagreeing/Agreeing, Asking for Opinions or Information, Challenging, Supporting, Modifying,

Table 2 Language function checklist

Types of functions	Language functions	Abbreviations
Informational	Giving personal info.	Giv
	Expressing opinions/preferences	Exp
	Elaborating	Ela
	Justifying opinions	Jus
	Comparing	Comp
	Speculating	Spe
	Describing a sequence of events/scene	Des
	Suggesting	Sug
Interactional	Disagree/Agreeing	Agr
	Modifying	Mod
	Staging	Sta
	Asking for opinions	Ask
	Commenting	Comm
	Greeting	Gre
	Negotiating meaning (check meaning, understanding, ask clarification, respond to required clarification)	Neg
Managing Interaction	Initiating	Ini
	Deciding	Dec

Persuading, Developing and Negotiating Meaning (He & Dai 2006). These functions were also reflected in the function list developed for the study.

To ensure the reliability of the coding, a second coder was involved, an experienced research assistant who was currently pursuing a PhD in applied linguistics at the time of the study. A portion of the data randomly selected from the full data set were coded by the second coder. A satisfactory level of agreement was achieved using Cohen's *Kappa* ($\kappa = 0.86$) on overall reliability.

Conversation analysis

The second research question sets out to identify the speaking features that distinguish tests rated at different levels. To answer this, Conversation Analysis was employed. The analysis is bottom-up and data driven and have been applied in existing literature to explore the relationship between candidate talk and scores (e.g., Lazaraton 1998, 2002; Seedhouse 2004). The candidate speaking features were defined as the features the candidates produce within a turn, which have the potential to lead to an increase or decrease in a candidate's score. For each of the features, extracts will be presented to demonstrate and compare to illustrate the typical differences between candidate performances.

Results and discussion

Language functions elicited in the test

As mentioned, a modified observation checklist was used for the analysis. The data confirmed that the types of function observed in each section of the test were consistent with the intended objectives of the test; basically covering the functions described in the SOPE standard (Renmin University of China, 2013a).

The average number of each function was calculated per candidate for the test. We can see from the Fig. 1 that among the language functions, the most frequently elicited one was Explanation, 3.28 on average for each candidate, followed up by Elaboration with the mean count of 2.11 per candidate and Justification of 1.56 per candidate. The mean numbers of the rest of functions are relatively small. The result is in accordance with the test specifications:candidates are expected to give clear description of given prompt (picture or charts) and express their opinions on the given topic.

The Figure also shows that in Section B of the test, the forms of interaction among candidates mainly were Disagreeing/Agreeing (0.81), Suggesting (0.50), Staging (0.44), Asking for opinion (0.72), Challenging (0.43), Supporting (0.46), Developing (0.61) and Greeting (0.70). Specifically, language to agree/disagree and developing personal opinions while asking for others' opinions were successfully observed. In the discussion, candidates usually stated their positions first, followed by justifying opinions. This part also started with candidates greeting each other and thanking for the interlocutor when they finished the task. Other forms of interaction, however, such as Modifying, Commenting and Persuading had fewer occurrences in this part of the test. How each function was realized were exemplified below, as the study should ensure that ways in which these functions were elicited were in line with the test designers' intentions.

Disagreeing/Agreeing

Excerpt 1:

> A: Do you know this is the job they dream of? I think most of us dream of a job with high salary and good social position. But I think no one is dreaming to be a person dealing with garbage, but our society really needs someone like that. So it's a problem of job hunting.…They don't think the job is of high quality.
>
> B: I agree with you, and I also think that university students should also take responsibilities. They may learn much about the books, learn how to write paper, but they don't learn in the university that how to apply their knowledge into real world. So they need more working experience
>
> C: I don't really agree with you. Because there is no major in the university like garbage, how to…school just don't teach students how to deal with garbage, but in college you can learn nearly whatever you want to learn. Maybe the school…there is no class…no courses in the school, but you can learn outside the school.

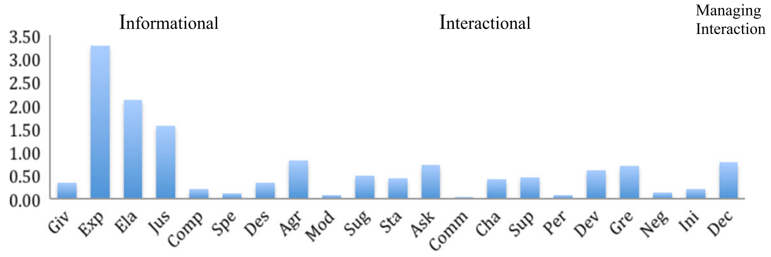

Fig. 1 Mean count of language functions elicited

The topic for discussion was "is studying abroad a worthwhile investment?" Excerpt 1 is one example of Disagreeing/Agreeing. Candidate B thought that university students should take social responsibility rather than adopt bias attitudes towards different jobs. He also believed that students needed more hands-on experience in college. However, Candidate C contradicted B that some knowledge can be learned outside university since university cannot cover all the skilled needed in the society.

Asking for opinions or information

Excerpt 2:

> C: But I think it should be discussed differently. I cannot agree with you, I must say. How do you think?

For excerpt 2, the topic for discussion was "True happiness lies in giving rather than taking." Candidate C expressed his disagreement with A, at the same time, he tried to ask for opinions of candidate B.

Challenging

Excerpt 3:

> C: It is very important you are happy at first. But when you give, when you help others, that happiness is spread among the people.
>
> A: Through giving, perhaps you can not only see the ability in helping but also see other people really get help from you.… How do you think actually giving really gives you happiness? How to explain this?

Here candidate C was discussing the importance of giving. But candidate A challenged her by asking a series of questions after analyzing both situations.

Supporting

Excerpt 4:

> C: Yes, I've heard about it. Actually, we know that the real world is more important than the virtual world, because the virtual world is not true. But don't you think that we rely more on the virtual world?
>
> B: Yeah, I think your point, that makes sense, because although nowadays we rely more on the virtual world, and the virtual world actually is doing us some good like it give us convenience. It just makes our real world more convenient to us.

In this excerpt, the topic was "The real world is more important than the virtual world". Candidate C acknowledged the importance of real world, but pointed out a common phenomenon that people nowadays heavily rely on virtual world. Candidate B supported his point by giving further evidence that the virtual world does bring convenience to people.

Modifying/Commenting

Excerpt 5:

> A: I also think that student today are not very interested or care about politics of our country. And I think in the future, we should care about politics more. For that purpose, we can do more things for our country. But I don't know how to do that.
>
> B: About you said, some students read fewer books; because they said they have no time, and they have a lot attractive do after class. And I think if you read the paper books at the library or you buy some books, it will waste you some time....

The topic for discussion was "How to promote after-class reading among college students?" Candidate A expressed her opinions in an unclear way; Candidate C gave her comments on Candidate A's idea and furthered the discussion to identify the incentives/motivations about reading e-books.

Persuading

Excerpt 6:

> A: We all know that in the teamwork, some kind of people they work very hard, they give their effort. Someone just stay lazy, and wait for others to give the results. But I think for those who work very hard, they make effort, may be very tired, but the pleasure they get from the procedure better than the comfort gain by those who don't work very hard.
>
> B: But we have to know this that some people only to give just like you said in teamwork. Some people they give because they want to make you feel like he is cleverer than you, he is stronger than you, and he is more powerful than you. So he choose to give not because he wants to make this thing better, just want to show that he is more powerful than you. That's a kind of abnormal kind of giving.
>
> C: So your idea is that giving means not happiness. He did that just for showing the power.

In this excerpt, the candidates were discussing around the topic "True happiness lies in giving rather than taking." Candidate A echoed ideas of Candidate C that some people work hard for their own pleasure. Candidate B then tried to persuade them to accept his view that there exits another explanation of the same phenomenon.

Developing

Excerpt 7:

> A: Yeah. I think so. As the old saying goes, when you give away a rose, the good smells will remain on your hands. I think that happiness will be pass on as B said. And I think that giving should be (re)commended in our society. And if everyone knows the meaning of giving, the society will be more harmonious. We all like a family. I think the world will be better.
>
> C: Yean, I agree. Talking about family, I think parents are a very good example. They always giving something to us, and never ask for return. …I think in the society, we should be like the family, like parents and child to give things to other people and not ask for return. Do you agree?

In Excerpt 7, Candidate A raised the point of building a harmonious society by passing on happiness. By concluding, he mentioned the concept of family. Building on this concept, Candidate C expanded and elaborated the role of parents as an example.

Thus, he also added another dimension of the topic that people in society, as one of the family members, should give and ask for no returns.

Negotiating meaning

Excerpt 8:

> C: ….And in my opinion, on the one hand we can give them the money, and love. But on the other hand, we can help them to develop their skills, to make them to learn something to make their own living. And I think the most important for the whole society is to develop everyone's skills to make our own living. It's very important.
>
> A: You mean that we must first do our best to make ourselves strong before we can help other gain the happiness. Is that right?

The topic of this round of discussion was "True happiness lies in giving rather than taking." Candidate C raised an important point during the discussion that people should develop life skills to make a living rather than just take money or receive help from others. In his turn, Candidate A clarified her opinion by asking a question.

As for the functions of managing interaction, some of the candidates initiated the conversation successfully (M = 0.21) often by greeting or asking others' opinions. Also some of the candidates even act as the role of the interlocutor to terminate and put an end to the discussion (M = 0.78). This also reflected a shared responsibility to maintain and developing the interaction, which is an important aspect reflected in the CEFR interpretation of spoken fluency.

Initiating

Excerpt 9:

> C: Well, the discussion topic well is about the happiness. Well, how do you think the giving and taking in this part?
>
> A: I think giving means giving help to others like give your care to others. But taking means maybe taking interest or other advantages from others.

At the beginning of the discussion, no one intended to start. Candidate C then initiated the discussion by repeating the topic and raised the questions to other two candidates.

j. Deciding

Excerpt 10:

> A: So, can we give a conclusion to our topic?
>
> B: And, yeah, so far we have talked about the same interest and the emotions we develop in our early time are all important factors about the friendship. And that's what makes our friendship last for a long time.
>
> C: And to add up, you forget the ambition.
>
> B: Yeah, the ambition.

The topic of this excerpt was "True friendship should stand the test of time". The three candidates all uttered their own opinions. After that, Candidate A initiated a conclusion for the whole discussion. Candidate B summarized the viewpoints raised by the three. Candidate C also contributed to adding up another point. Finally the three candidates reached the agreement regarding the topic, successfully completing the discussion task.

Besides analyzing and exemplifying the individual function elicited, the study also examined whether the main functions observed in the data are consistent with the expectations of the test. In doing so, the researchers also related those functions to three main categories of language functions (*informational, interactional* and *managing interaction*) in order to obtain an overall picture of student performance.

As shown in Fig. 2, the second part of the test mainly elicited *informational* functions such as Giving personal information, Expressing opinions, and Elaborating. It accounted for 57% of the total number of functions, followed by some *interactional* functions like greeting and thanking, 36%. *Managing interaction* functions came as the least elicited type of functions, accounting for 7% of the total number of functions.

While the majority of functions were observed at a range of levels, some mainly appeared at higher proficiency levels. For example, functions like speculating, staging appeared at higher levels. A major difference among higher and lower proficiency levels can be found in the *interactional* and *managing interaction* functions. According to the test design, candidates at higher levels are expected to engage in functions such as initiating a topic, developing/supporting it, or finding opportunities of change the topic. These gradation of the functions are closely linked to the cognitive demands of the tasks (Bygate 1987).

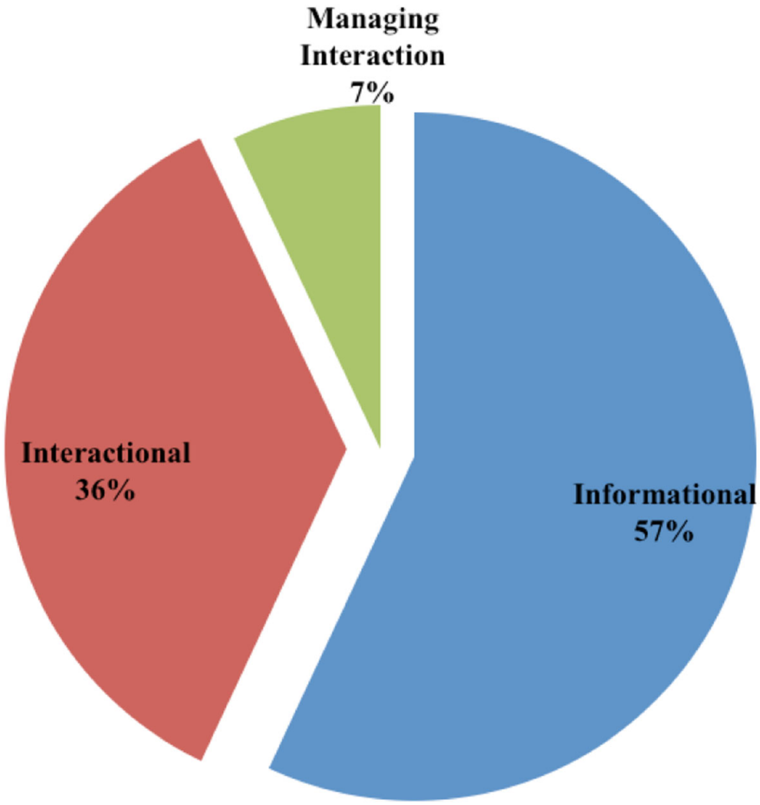

Fig. 2 General category of language functions elicited

The results also showed low frequency of occurrence of language functions such as Modifying, Commenting, Persuading, Negotiating of Meaning, though these functions have been clear stated in the SOPE Standards and test specifications. The phenomenon may be due to several possible reasons. First, candidates might lack of confidence in their ability to handle the group discussion task as communicative interaction. They focused more on their own linguistic accuracy rather than construct and maintain the interaction with other candidates. Or they would rather agree with others' opinions than continue their own lines of thoughts. Second, in testing situations, candidates might interpret the discussion task as a test item rather than an authentic discussion in real life. In other words, there exist a mismatch between what learner *can* do and what they *do* in testing situations. They may concentrate on their own performance rather than respond to others' thoughts actively, ignoring the interactive nature of the task.

Speaking features that distinguish proficiency at each level

The second research question focuses on individual speaking features exhibited during the delivery of a candidate's turn. The following speaking features were then identified based on the qualitative analysis of the spoken data.

Length of turn

It was found that candidates at a higher level tended to have more instances of extended speech turns to develop the topic.

Extract 11

A: So what do you think that makes friendship (10 s)? Maybe you say that some factors may influence our long term relationship.

C: I think…I also have a friend that our friendship last for nearly ten years. Yes, ten years. Although I'm still eighteen years old, but it last for ten years. My friend has the same hobby that (5 s). I have concerned that we share the same hobbies that we share the same ambition like we will going to Beijing and we want to work here. And that's the same ambition we will work, we will find more. I think why we are still keeping this relationship? It might because the same hobby, and the same ambition. Although she used to study in Hunan province, but in my senior high school, I used to study in Beijing. Though the distance is so far, we can still write letters to keep in contact. I think this is very important to let our friendship stand the challenge of the time.

A: Yes, but can this happen that maybe after a long time that we are all change, and things may be not just like the original thing. Maybe our personalities, our interest.

In this extract, it can be seen that candidate C produced relatively a lengthy turn, which was closely related to his score (Band 5). Also, there is some evidence that candidates at lower levels produced short turns with lengthy pauses (longer than 10 s).

Choice of words and syntax

Candidates with a high score might develop a discussion topic using lexical items which are less common and this portrayed them as having higher level of cognitive processing. Also, the candidates at a higher level displayed the ability to construct a sentence with more complex sentence patterns.

Extract 12

> A: So how can you get rid of it? How can you stay away from the cell phone, and resist the temptation of it? Many of us cannot do it, just like me. I have got addicted to it.
>
> C: Sometimes I don't take my cell phone, just put it in my dormitory. And I go out without it.
>
> A: Yeah, so how can distract our attention to the virtual world and focus our attention on the real world? So what can we do?

In this extract, the candidate demonstrated a number of speaking features that may positively impact his score. First, he asked questions to other candidates, which is a way for eliciting others' opinions and initiating the discussion. In this regard, he acted as the role of facilitator in this discussion; in doing so, he positioned himself as a high achiever. He also employed a number of less commonly used vocabulary related to the topic compared with other candidates, e.g., "temptation", and "addict", "virtual".

Extract 13

> B: Yeah, I feel like the exchange of giving and taking not only happen in families, not only the family members share this. Did you hear the news that such a murdering event that happened in Zhaoyuan? Everyone just stood by and turned a cold eye on it, and didn't offer their hand. And that young lady was killed to death. So if someone lend their hand to her, all of this kind of evil person will be under the control, and the life wouldn't be gone. And happiness and harmony of the society will be continued.

The candidate in the above extract (Band 5) successfully developed the topic in a meaningful way by using examples and further raised his points (e.g., "Did you hear the news that…"). In addition, he constructed different sentence patterns, used prefabricated chunks within a turn (e.g., "turn a cold eye", "lend their hands"), and several sentences using passive voice.

Hesitation markers

Another feature that shows patterns within candidate performance is hesitation marker (Ellis & Barkhuzen 2005), such as *er* and *erm*. As a regular feature of ordinary conversation, high concentrations of hesitation markers will disrupt the flow of speaking production. It is more common to find higher concentrations of these features in some of the candidates in lower band levels. In the following extract, the candidate performance (Band 3), in above extract, is replete with examples of hesitation markers.

Extract 14

> My topic is studying abroad, learning about the (for…learning…er…learn about the foreign…er…foreign culture. We all knows foreign culture is very colorful, especially American culture. Er…so it's…er…so in this world, er…er we…er…er learning…er learning about foreign culture is our duty to…deal with this…er…this…this (laughter) world…in this world.

Topic coherence

Candidates at higher band levels usually developed the topic coherently, using cohesive devices such as "also", "in addition" to connect clauses. Comparatively, candidates with lower scores sometimes struggled to construct an argument, as in the extract below. Turns of this group of candidates often featured lengthy pauses.

Extract 15

> B: I don't think so. I think Internet is good. On the Internet we can also send other through QQ. We can talk some topics in…er…er…like in the real life. I think Internet can give us more benefits than…er…er…yes give us more benefits.

From another perspective, it was also found that for every identified pattern, a significant number of counter cases were identified that contradicted this pattern. There was therefore no individual speaking features that can be said to rigorously distinguish among various levels. Though qualitative, the overall picture is similar to some quantitative investigations in the literature. Just as Brown (2006:71) concluded "Overall, the findings indicate that while all the measures relating to one scale contribute in some way to the assessment on that scale, no one measures drives the rating; rating a range of performance features contribute to the overall impression of the candidate's proficiency".

Conclusions

The present study investigated the validity of a university-based speaking assessment in mainland China. The degree of interaction among candidates in the group discussion was analyzed in terms of language functions included in the test syllabus, developed based on the CEFR. Quantitative analysis revealed that the majority of language functions intended by the test syllabus were observed at a range of levels in candidate performances.

Further qualitative investigations identified some speaking features that have the potential to affect scores of the candidate, including: lengthy of turn, choice of words and syntax, hesitation markers and topic coherence. Besides, complex language associated with higher levels was used to address abstract subject matter. This progression means that cognitive and linguistic complexity are both required to carry out tasks at the higher levels.

To develop language tests, CEFR provides a systematic framework and point of reference for developing and comparing tests in different educational and assessment contexts. In the current case, it has been helpful in operationalizing the construct and defining the level descriptors. Since the empirical validation stage for linking the TOPE to the CEFR has not completed yet, whether the latter is effective to objectively compare students' speaking proficiency across levels is unknown. Through the overall process of development and implantation of the TOPE, there were also difficulties. First, students in Renmin University are relatively high proficient language learners; the project team met difficulties to differentiate between the abilities of highly proficient learners just based on the descriptors of CEFR. Second, there lacked a comparability of the test to other standard language assessments in China. In this regard, the need for a local language framework that can act as a point of reference and as an indication of language achievements in schools and universities is therefore urgent, as proposed by Jin *et al* (2016).

For practical applications, developing and implementing a university-based assessment has entailed a major change in the teaching and learning and the provision of university supports. Consider teaching and learning first, teaching materials and methods have been calibrated to the features identified. Explicit exercises have also been designed to target these features at different levels, which can make the teaching more effective. For example, teachers in the university have provided instructional scaffolding to support student learning e.g., how they develop speaking tasks, how students learn from playing an active role in the classroom. It is also worth paying attention to the warning that it is important to avoid assessment practices becoming mechanistic, ritualized and ultimately meaningless and boring to students (James 2011).

Yet it must be noted that the research method has its limitations. There is no clear-cut boundary for each language function and some spoken discourse involves a variety of language functions. Also, a checklist of speaking features is not the whole story in assessment, learning and teaching. The language functions can be realized in a variety of ways. Future research into functional resources would be beneficial to inform not only contextual aspects of the TOPE but also key features relating to its scoring validity, e.g., assessment criteria and rater training. At the same time, it cannot deny that this approach can provide insights into what really goes on in the speaking tasks, which is difficult to access by other means. Future study can therefore be conducted to identify clusters of features at different levels that underpin each proficiency level, which will give use insights into the time course of language learning and teaching.

Funding
This paper was based on the project "The Applicability of the CEFR for English Language Education in China" (GZ20140100) funded by Foreign Language Teaching and Research Press.

Authors' contributions
LL carried out the main study and performed the data analysis. GJ conceived of the study and participated in the data collection and coordination of the study. Both authors read and approved the final manuscript.

Competing interest
The authors declare that they have no competing interests.

Received: 29 October 2016 Accepted: 31 January 2017
Published online: 06 February 2017

References

Atkinson, J. M., & Heritage, J. (Eds.). (1984). *Structures of social action: studies in conversation analysis*. Cambridge: Cambridge University Press.

Brooks, L. (2003). Converting an observation checklist for use with the IELTS speaking test. *Cambridge ESOL Research Notes, 11*, 20–21.

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. *IELTS Research Reports, 6*, 71–89.

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

Cheng, L., & Curtis, A. (Eds.). (2010). *English language assessment and the Chinese learner*. New York: Routledge, Taylor and Francis Group.

Council of Europe. (2001). *The common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Fan, J., & Ji, P. (2013). Examining the validity of the Fudan English Test: test data analysis. *Foreign Language Testing and Teaching, 2*, 45–53.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23*(3), 370–401.

Jacoby, S., & Ochs, E. (1995). Co-construction: an introduction. *Research on Language and Social Interaction, 28*(3), 171–183.

James, M. (2011). Assessment for learning: research and policy in the (Dis)United Kingdom. In R. Berry & B. Adamson (Eds.), *Assessment reform in education: policy and practice* (pp. 15–32). Dordrecht: Springer.

Jin, T., Wang, Y., Song, C., & Guo, S. (2008). An empirical study of fuzzy scoring methods for speaking tests. *Modern Foreign Languages, 31*(2), 157–164.

Jin, Y., Wu, Z. M., Alderson, C., & Song, W. (2016). Developing a framework of reference for English language education in China: challenges at macro- and micro-political levels. *Language Testing in Asia, x*, xx.

Lazaraton, A. (1998). *An analysis of differences in linguistic features of candidates at different levels of the IELTS Speaking Test* [Unpublished study commissioned by UCLES].

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Mendelsohn, D. J. (1989). Testing should reflect teaching. *TESL Canada Journal, 7*(1), 95–108.

Ministry of Education. (2007). *College English Curriculum Requirements*. Shanghai: Shanghai Foreign Language Education Press.

O'Sullivan, B., Weir, C., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33–56.

Pan, Q. (2013). An empirical study on the correlation between fluency and accuracy in Chinese EFL learners' oral production and its longitudinal trends between different proficiency levels. *Foreign Language Research, 4*, 100–106.

Renmin University of China. (2013a). *Standards for Oral Proficiency in English*. Beijing: Renmin University Press.

Renmin University of China. (2013b). *Test for oral proficiency in English: test syllabus*. Beijing: Renmin University Press.

Seedhouse, P. (2004). *The Interactional Architecture of the Language Classroom: a conversation analysis perspective*. Malden: Wiley-Blackwell.

Seedhouse, P., Harris, A., Naeb, R., & Üstünel, E. (2014). The relationship between speaking features and band descriptors: a mixed methods study. *IELTS Research Reports Online Series, 2*, 1–30.

Sun, H., & Wei, M. (2012). Modern measurement analysis of an oral rating scale. *Foreign Languages and Their Teaching, 6*, 66–70.

Tang, L. (2014). Features of conversation interaction in pared speaking test. *Foreign Languages and Their Teaching, 5*, 36–41.

TOPE Project Group. (2013). *Unpublished report for developing Test of Oral Proficiency in English*. Beijing: Renmin University of China.

Tsinghua English Proficiency Test Group. (2012). *Tsinghua English Proficiency Test Syllabus* (2nd ed.). Beijing: Tsinghua University Press.

Wang, Y. (2015). Distinguishing features in second language speaking assessment. *Journal of PLA University of Foreign Languages, 38*(2), 102–108.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Yi, Q., & Zeng, L. (2004). Scoring pattern and its application to the evaluation of oral language skills. *Modern Foreign Languages, 27*(1), 81–86.

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45*(1), 3–42.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com