TEST REVIEW

College English Test–Spoken English Test (CET-SET)

Overview

The College English Test (CET) is a national standardized English language test for college and university students in China, which has been in operation since 1987 as an assessment tool to see whether these students have met the curriculum requirements of their compulsory College English course as specified in the College English Teaching Syllabus (State Education Commission, 1985, 1986; Higher Education Department of the Ministry of Education, 1999) and College English Curriculum Requirements (Higher Education Department of the Ministry of Education, 2007). The CET measures the four major language skills separately, i.e., listening, reading, and writing in the CET and speaking in the CET Spoken English Test (CET-SET). The CET-SET, which is the focus of this review, currently has an annual test population of over one million (before the pandemic).

Test purpose

The purpose of the CET-SET is to measure the oral English proficiency of college and university students and to examine whether these students meet the required levels of oral English language abilities as specified in their compulsory College English course, thus making the CET-SET a curriculum-based achievement test. Since the four-year College English course is divided into the foundation stage (the first two years) and the post-foundation stage, the CET-SET correspondingly consists of CET-SET Band 4 targeted at students who have completed their English learning at the foundation stage, and CET-SET Band 6 targeted at students who have completed post-foundation studies.

Length and administration

The CET-SET is designed and developed by the National College English Testing Committee of China (NCETC). It is administered by the National Education Examinations Authority (NEEA), an institution directly under the supervision of the Ministry of Education and appointed by the Ministry to undertake educational examinations. The CET-SET is currently held twice a year in May and November at authorised CET-SET test centres. As an optional component for students, this speaking test is administered separately from the CET. The total test time is about 15 minutes for CET-SET Band 4 and approximately 18 minutes for CET-SET Band 6.

Scores

The test results of the CET-SET are reported on a graded score scale, i.e., A, B, C and D, with A indicating the highest level and D a fail. The final graded score is converted from the average of the total raw scores awarded by two authorized CET-SET markers.

Author and publisher

The National College English Testing Committee is responsible for test design and development. The National Education Examinations Authority is in charge of test management and implementation.

Information available

The CET official website (http://cet.neea.edu.cn/) contains information about the test content and format of the CET-SET, the oral skills to be measured, assessment criteria, scoring methods, score reporting systems as well as sample test papers. Video clips introducing the testing procedure are also available on

the website to help test takers familiarise themselves with the test and understand what they are expected to do in the test.

Price

The prices are 50 RMB for both CET-SET Band 4 and Band 6 (at the time of writing, the exchange rate of USD vs. RMB was roughly 1:7).

General description

Development background

In the late 1980s when the CET was inaugurated, speaking was not an essential requirement for college students and was therefore not included in the test (Jin, 2020). With China's further opening up, the CET-SET was introduced in 1999 to meet the increasing need for oral communication in English. The original CET-SET consisted of only one level, targeted at students who had completed their English learning at the post-foundation stage.

The test adopted a face-to-face interview format, with a view to 'encouraging students to participate more actively in interactive communication' (Jin, 2010, p. 52). In the test, three test takers and two examiners formed a test group to complete three monologic tasks and a peer-to-peer discussion task. Since its first administration, the CET-SET has received an increasing amount of attention from college students and teachers in China (Jin & Chen, 2002). By 2012, a total of 58 CET-SET test centres had been established and more than a thousand examiners had been trained (Jin, 2022). However, the scale of the face-to-face test was seriously constrained by the number of qualified examiners as well as the logistics required for test implementation. For this reason, only those who had reached a designated minimum score on the CET written tests were eligible to take the speaking test.

With the intention of improving test efficiency and practicality (Jin & Zhang, 2016), the CET-SET transitioned from the face-to-face format to a computer-based delivery mode in 2013. The minimum CET score requirement has been removed since then. The computer-based CET-SET adopted a paired format, in which two test takers form a pair to complete a series of oral tasks, including a paired discussion task. Later, the CET Committee also developed a lower-level computerised speaking test, CET-SET Band 4, to accommodate the needs of students who have completed foundation stage studies. The prior computer-based test then became known as CET-SET Band 6 and remained unchanged in terms of test content.

Test format

A group interview format, as previously mentioned, had been used in the face-to-face CET-SET before 2013, where two examiners interviewed three test takers at a time, to maximize assessment efficiency. In the transition from the face-to-face mode to the computer-based mode, a peer-to-peer discussion task was still incorporated into the test, but a paired format was used instead. The primary reason for the switch was the concern with the difficulty in distinguishing more than two test takers' voices in the audio recordings when scoring the discussion task in the computer-delivered mode (Zhang, 2019). In the current version of the CET-SET, therefore, two test takers are randomly paired by the computer system and engage in three monologic tasks and a peer-to-peer discussion task. Test takers' oral responses are automatically recorded by the computer and are scored by raters after the test. To minimise interlocutor effects, pre-recorded videos of human (native speaker) examiners are employed to play the role of interlocutors in the relevant monologic tasks. Throughout the computer-based test, test takers can 'see' each other via photos on the screen; live video is not used primarily due to a concern about bandwidth. When engaging in paired discussion, the test takers talk to each other via headphones,

thus interacting in a non-face-to-face manner.

Test structure

Both bands of the CET-SET consist of four tasks. At each level, all tasks in a test relate to one broad topic. Table 1 provides an overview of the test content and format of the CET-SET. For the CET-SET Band 4, Task 1 asks test takers to read aloud a text of approximately 120 words. Task 2 requires test takers to answer two questions. The first question is closely linked to the content of the read-aloud text (Task 1), while the second question is relevant to the theme of the text. Task 3 is an individual presentation task, in which test takers give a speech based on the given visual prompt (pictures, diagrams, etc.). Task 4 is an interactive task which engages test takers in a paired discussion.

Table 1. Test content and format of the CET-SET

Test	Part	Task format	Pattern of interaction	Preparation /response time	Length
CET-SET Band 4	Task 1	Reading aloud	Monologic	45/60 seconds	About 8 minutes
	Task 2	Question and answer	Monologic	0/40 seconds	
	Task 3	Individual presentation	Monologic	45/60 seconds	
	Task 4	Paired discussion	Interactive	1/3 minutes	
CET-SET Band 6	Task 1	Warm-up questions	Monologic	0/30 seconds	About 7 minutes
	Task 2	Individual presentation	Monologic	1/1.5 minutes	
	Task 3	Paired discussion	Interactive	0/3 minutes	
	Task 4	Further-check questions	Monologic	0/45 seconds	

For the CET-SET Band 6, Task 1 asks test takers to answer a short question, which is intended to help test takers warm up at the very beginning of the test. Task 2 requires the paired test takers to give, in turn, an individual presentation

based on related prompts. Task 3 is a discussion requiring the paired test takers to exchange opinions about a topic related to the Task 2 prompt. The last task asks each test taker to answer one more question, which is intended as a further check on the test taker's oral proficiency. For a sample of the test, see National College English Testing Committee (2016).

One of the major differences between the CET-SET and other large-scale speaking assessments is that all tasks in any given test version revolve around a particular topic. Topics used in the CET-SET Band 4, the lower level of the test, are mostly of a concrete or personal nature such as personal interests, future work and so on. Popular topics for the CET-SET Band 6, the higher level of the test, are of a more abstract nature such as social issues, cultural differences, and so on.

As shown in Table 1, both the CET-SET Band 4 and Band 6 include a peer-peer discussion task, which is a unique feature of the test as compared to other computer-based speaking assessments and will be discussed further in the speaking construct section. The discussion task in the CET-SET Band 4 encourages test takers to reach an agreement through cooperative discussion. For instance, test takers may be asked to work out a plan to visit a museum by discussing which museum to visit, schedule of the visit and means of transportation. The prompts are mostly pictures, tables or charts, and test takers have one minute of preparation time. The discussion task in the CET-SET Band 6 instructs test takers to argue and debate on a topic with some degree of controversy. For example, test takers may give an individual presentation on the topic of artificial intelligence, with foci on the progress in AI technology and its impacts. Following the presentations, they engage in a discussion of whether AI will render many people jobless. The prompts are mostly in the form of text, and no planning time is allowed for the discussion task at this level.

Scoring

Each test taker's performance is double scored by raters who are English language teachers at the tertiary level with an advanced level of English language proficiency. Analytical scores on three criteria, accuracy and range, size and discourse management, and flexibility and appropriacy, are given by each rater. The descriptors for each set of criteria are provided in Table 2. With their focus on the ability to communicate effectively, the CET-SET assessment criteria address the accuracy, fluency, and appropriacy of candidate speech in relation to the various subcategories of linguistic competence, discourse competence, and pragmatic competence (Jin, 2000a, cited in Zhang & Elder, 2009). The raters each score independently using a rating scale ranging from 1 to 5 for each of the three analytic rating categories, with scores assigned based on a cumulative impression of the test taker's speaking performance across all tasks (Yang, 1999).

Table 2. The CET-SET assessment criteria

	Category	Descriptors
1	Accuracy	Accuracy in pronunciation, intonation and use of grammar and vocabulary
	Range	Complexity and range of vocabulary and grammatical structures
2	Size	Size of contribution made by the candidate
	Discourse management	Ability to produce extended and coherent discourse
3	Flexibility	Flexibility in dealing with different situations and topics
	Appropriacy	Appropriacy in the use of linguistic resources according to communicative contexts

Note. This translation is adapted from Zheng and Cheng (2008).

The only exception is the read-aloud task in the CET-SET Band 4, which is automatically scored based on a rating scale ranging from 1 to 5. The scoring criteria for the read-aloud task include: (1) accuracy: appropriateness of pauses between meaningful chunks and accuracy in pronunciation and intonation; (2) fluency: fluidity of expression and frequency of self-repetition and self-correction; and (3) content coverage: how much of the text was read out (see

National College English Testing Committee, 2016 for the read-aloud rating scale). The total score of Band 4 is a combination of two ratings: one given by the automated scoring system for the read-aloud task, and one awarded by the human raters (averaged over two raters, with disagreement necessitating a third rater) for all the other tasks.

The test results of the CET-SET are reported on a graded score scale, i.e. A, B, C and D. Test takers who achieve a grade of C or above receive a score report which indicates the grades (A, B or C) and provides detailed grade descriptions. For instance, Band 4 grade descriptions focus on the test taker's ability to participate in oral interaction in English on familiar topics, express personal views about familiar issues, and narrate or describe general events and phenomena. As Band 4 and Band 6 differ in speaking ability requirements and difficulty levels, different grade descriptions are used for the two bands accordingly (see National College English Testing Committee, 2016).

Strengths and weaknesses

Speaking construct

Reciprocal interaction has long been considered as 'integral to the construct of oral proficiency and not simply an optional component' (O'Loughlin, 2001, p. 169). Hence, excluding tasks that tap into the interactional element of the speaking construct may lead to potential construct under-representation (Galaczi, 2010; Nakatsuhara et al., 2017). When the CET-SET was developed, interactional competence was viewed as an essential component of the speaking construct (Jin, 2020). Based on a national survey of the needs for oral communication in English prior to the development of the CET-SET (Huang, 1999), the peer-to-peer discussion task was adopted to better represent language use in the educational domain in mainland China. Talking to peers

could also avoid the issue of imbalanced power relationship between the examiner and the test taker in an oral interview (van Lier, 1989; Weir, 2005) and elicit more natural and dynamic conversations (Fan & Yan, 2020; Zhang, 2019).

When the CET-SET went online, paired testing was adopted to retain the peer-peer interaction task so that assessment of interactional competence could still be achieved in a computer-based delivery mode. The paired format adopted by the CET-SET, which elicits not only oral production but also reciprocal interaction, is unique in that human interaction has not featured in other large-scale computer-based speaking tests (Zhang & Jin, 2021). Jin (2019) noted that the computer-mediated discussion task in the CET-SET also has the advantage of simulating the increasingly popular mode of oral communication such as talking via Skype or WeChat, participating in online discussion, and attending online conferences.

However, the computer-based CET-SET assesses audio-only interaction, which may differ from online interaction in which participants can see each other via video cameras. For instance, in investigating the impact of test mode, face-to-face versus computer-based, on test takers' use of communication strategies in the discussion task of the CET-SET, Zhang (2019) found that some differences emerged in the use of some cooperative strategies and meaning-negotiation strategies between the two delivery modes, although most communication strategies were used equally across the two conditions. Her research also revealed the increased challenge for turn-taking management when test takers communicated without visual cues, which somewhat affected the smooth flow of oral interaction. In addition, as much literature (e.g., Iwashita, 1996; Norton, 2005) has pointed to the issue of interlocutor effects in paired oral tasks, another important direction for future research is to look into whether interlocutor characteristics, such as gender, personality, proficiency level and

style of interaction, have considerable impact on test-taker performance and test results, thus impinging on the validity of the computer-mediated discussion task.

Scoring reliability

Over the years the CET Committee has established a qualified team of markers and a strict system of marking quality control (Jin & Yang, 2018). So far, a total of 58 CET-SET test centres have been set up and over a thousand raters have been trained. After each administration of the test, the CET Committee decides on the benchmark responses (representative performances for each level) and sample responses to be used for rater training. For every test administration, all raters need to have successfully completed rater training sessions conducted at the CET-SET test centres before proceeding with their rating work. The training sessions follow the standard procedures: 1) studying the rating scales, 2) scoring the training recordings to try out the scoring procedure, and 3) reaching a consensus score and comparing it with the benchmark (Zhang & Elder, 2009).

To maintain the marking quality, the practice of double scoring has remained unchanged either in the face-to-face format or in the computer-delivered mode, although the scoring process is rather labour-intensive and time-consuming for a speaking test of such a scale. When the CET-SET shifted to a computer-based format, an online marking system was developed by the CET Committee to further enhance the efficiency and quality of marking. The marking system enables real-time monitoring of rating performance and provision of immediate feedback to raters, which help improve inter- and intra-rater consistencies. However, information about rater reliability has not been published so far. In the interests of greater transparency, as Zhang and Elder (2009) pointed out, it would be helpful if CET-SET administrators could

consider making such information available to the test users.

Use of automated scoring

To cope with the increasingly heavy workload of human scoring, the CET Committee, in collaboration with an intelligent speech and artificial intelligence company, developed the CET-SET automated scoring system with a view to enhancing scoring efficiency. As mentioned earlier, only the read-aloud task in Band 4 is currently scored by this system. Although machine scoring for the open-ended tasks is still under trial stage, there has been evidence suggesting that its consistency with human scoring falls into a reasonable range. Adopting Knoch and Chapelle's (2018) argument-based approach to validating rating practices, Jin et al. (2020) conduced a preliminary validation of the scoring system, with their focus on three inferences to be made from the scores generated by the machine: the evaluation, generalisation and explanation inferences. Through analysing the human and automated scores of two test forms used in a live CET-SET Band 4 test, the researchers found a sufficiently high accuracy rate of speech recognition (over 98% for the read-aloud task and over 95% for the open-ended tasks) and a reasonably high Pearson correlation (0.83 and 0.85) between the human and machine scores. For both test forms, about 75% of test takers were assigned the same grade by the machine and human raters, although human raters made finer distinctions at the high and low ends of the proficiency levels. Experts' judgments of the criterial features that could distinguish between proficiency levels based on the machine scores revealed that the system was more sensitive to features of linguistic accuracy and content relevance and richness than to features such as pronunciation and intonation and use of communication strategies.

Currently, research is being conducted to examine the impact of rating methods on automated scoring of open-ended tasks (X. Zhang, personal communication,

November 22, 2022). Preliminary findings suggest that the CET-SET automated scoring system can achieve a reasonably high degree of consistency with human scoring when adopting a task-based analytic rating method (assigning task scores analytically). Similar results were produced when a task-based holistic rating method (assigning task scores holistically) was used. And the agreement rates achieved on Band 4 and Band 6 were quite close. Nonetheless, human raters were again found to be better at distinguishing high- and low-end test takers, which is consistent with Jin et al. (2020)'s findings. Although these studies provide some evidence of the CET-SET automated scoring system's reliability in scoring open-ended tasks, clear interpretation of the scores generated by the system is still quite challenging. As Jin et al. (2020) noted, it is essential that test developers make sure the machine scores are meaningful and informative to test users before being used in operational tests as the second or sole rater.

Washback impact

The implementation of the CET-SET is intended to bring about positive washback effect on the teaching and learning of English in universities in China (He & Dai, 2006; Yang, 1999). The development of the test over the past two decades has provided solid evidence of the powerful impact the test has on teaching and learning. Although the CET-SET is an optional test, today the number of students who register for the test has exceeded one million a year, indicating that 'great attention has been drawn to the teaching of spoken English at the tertiary level in China' (Jin & Yang, 2006).

Chen and Tao (2001) noted that the implementation of the CET-SET was important in that students could be motivated to practice their spoken English when preparing for the test and this could lead to improvement in their oral English proficiency. Furthermore, the test has been shown to have a positive

effect on college students' attitudes toward English learning (Tang & Peng, 2004). Similarly, Jin (2000b) found that after the CET-SET was introduced, students showed greater interest in learning spoken English both inside and outside the classroom, especially when their university became one of the CET-SET test centres. She further noted that many universities attached greater importance to enhancing students' language use ability and some of them even developed CET-SET related teaching materials to help students improve their speaking ability. However, research in this regard is still quite limited and prior research mostly focused on the face-to-face CET-SET. Given the switch to the computer-based CET-SET in recent years, further empirical research on the washback impact of the new test format is urgently needed.

Summary

The introduction of the CET-SET is undeniably a strong supplement to the CET test battery, as speaking is an essential component of communicative competence. The CET-SET is designed and developed in accordance with the curriculum requirements of the College English course. Since it was launched in the late 1990s, the speaking test has played an active and positive role in promoting college and university students' communicative language ability (Jin & Yang, 2006). The initial adoption of the most direct assessment format, a face-to-face oral interview, and the use of a paired format after the test transitioned to a computer-delivered mode are evidence of the endeavours made by the CET Committee to prioritise the construct of oral interaction and promote positive effects of the test on the teaching and learning of spoken English. To accommodate the growing number of test takers, more effort will need to be put into addressing important issues such as the availability of facilities as well as the number of qualified raters. As the operational test comprises multiple test sessions due to the large number of test takers, the

increasing volume of test sessions will also necessitate the development of a larger pool of test items (Jin, 2019).

Reviewed by Lin Zhang

Shanghai Jiao Tong University

References

Chen, J., & Tao, W. (2001). CET-Spoken English Test and its washback effect. *Shandong Foreign Languages Teaching*, *82*, 82–84.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, *11*:330.

Galaczi, E. D. (2010). *Face-to-face and computer-based assessment of speaking: Challenges and opportunities*. Retrieved June 25, 2016, from https://www.researchgate.net/publication/281090096

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

Higher Education Department of the Ministry of Education. (1999). *College English Teaching Syllabus (revised version)*. Shanghai Foreign Language Education Press.

Higher Education Department of the Ministry of Education. (2007). *College English curriculum requirements*. Shanghai Foreign Language Education Press.

Huang, S. H. (1999). *The development and validation of the College English Test–Spoken English Test (CET-SET)*. Unpublished doctoral dissertation, Shanghai Jiao Tong University.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing,*

5(2), 51–66.

Jin, Y. (2000a). *Feedback on the CET Spoken English Test and its backwash effect on the teaching of oral English in China*. Paper presented at the third International Conference on English Language Testing in Asia, Hong Kong.

Jin, Y. (2000b). The washback effects of College English Test–Spoken English Test on teaching. *Foreign Language World*, *80*(4), 56–61.

Jin, Y. (2010). The National College English Testing Committee. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 44–59). New York & London: Routledge, Taylor & Francis Group.

Jin, Y. (2019). Testing tertiary-level English language learners: The College English Test in China, in Su, L. I-W, Weir, C. J., & Wu, J. R. W. (eds). *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts* (pp. 101–130). London/New York: Routledge.

Jin, Y. (2020). Context validity in language assessment: test operations and conditions for construct operationalization. In Taylor, L. & Saville, N. (eds.) *Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)* (pp. 83–104). Cambridge: Cambridge University Press.

Jin, Y. (2022, March). *Decision making in language testing: Intersections of policy, practice, and research*. Paper presented at the 43rd Language Testing Research Colloquium, Online.

Jin, Y., & Chen, L. (2002). *Looking into the process of taking the CET Spoken English Test*. Paper presented at the International Language Testing Conference, Shanghai.

Jin, Y, Wang, W, Zhang, X., & Zhao, Y. (2020). A Preliminary Investigation of the Scoring Validity of the CET-SET Automated Scoring System. *China Examinations, 7*: 25–33.

Jin, Y. & Yang, H. (2006). The English proficiency of college and university students in China: As Reflected in the CET. *Language, Culture and Curriculum, 19* (1): 21–36.

Jin, Y. & Yang, H. (2018). Developing language tests with Chinese characteristics: Implications from three decades of the CET. *Foreign Language World, 2*: 29–39.

Jin, Y. & Zhang, L. (2016). The impact of test mode on the use of communication strategies in paired discussion. In Yu, G. & Jin, Y. (ed.) *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums* (pp. 61–84). *Palgrave Macmillan.*

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4): 477–499.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, *14*(1), 1–18. https://doi.org/10.1080/15434303.2016.1263637

National College English Testing Committee. (2016). *College English Test syllabus (revised version)*. Shanghai Jiao Tong University Press.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal, 59*(4), 287–297.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests: Vol. 13. Studies in language testing*. Cambridge University Press.

State Education Commission. (1985). *College English teaching syllabus (for college and university students of science and technology)*. Shanghai Foreign Language Education Press.

State Education Commission. (1986). *College English teaching syllabus (for college and university students of arts and science)*. Shanghai Foreign Language Education Press.

Tang, Y., & Peng, J. (2004). Washback effect of the CET–SET on English learning. *Foreign Language World*, *99*, 25–30.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23*(3), 489-508.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Yang, H. (1999). The CET–SET design principles. *Foreign Language World*, *75*, 48–57.

Zhang, L. (2019). *The Impact of Test Mode on the Use of Communication Strategies: The Case of CET-SET Paired Discussion*. Unpublished doctoral dissertation, Shanghai Jiao Tong University.

Zhang, L., & Jin, Y. (2021). Assessing interactional competence in the computer-based CET-SET: An investigation of the use of communication strategies. *Assessment in Education: Principles, Policy & Practice, 28*(4), 389-410.

Zhang, Y., & Elder, C. (2009). Measuring the Speaking Proficiency of Advanced EFL Learners in China: The CET–SET Solution. *Language Assessment Quarterly*, *6*(4), 298-314.

Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Language Testing*, *25*(3), 408–417.