TEST REVIEW

The listening and speaking test of NMET Shanghai

The focus of this review is the listening and speaking test ("the test" hereinafter) of the National Matriculation English Test in Shanghai (NMET Shanghai), one of first listening and speaking tests targeting candidates for tertiary education in Mainland China. The aim of the test was to change the situation of "dumb English" among Chinese learners, who may read and write proficiently but tended to fail in oral communication. As such, the test has achieved its intended positive washback effect to some extent. This review first introduces the history of this test and its structure, then analyzes its validity, reliability, and impact.

General description

Test development background

China's National College Entrance Examination, also widely known as *Gaokao,* is high-stakes and therefore a major event for the whole nation. In 2022 alone, there were 11.93 million candidates, according to China's Ministry of Education (MOE). *Gaokao* is in fact the general name of a series of tests for various subjects, with Chinese language, mathematics and foreign language (mainly English) being the most important, as they each make up about one-fourth of a candidate's total *Gaokao* score.

Considering the imbalance of resources and quality of education, China's *Gaokao* has different versions in different provinces and municipalities. Shanghai, being one of the first ports of China to be opened to Western countries, has been authorized by the MOE to develop and administer a localized English *Gaokao* test since 1985, officially recognized as NMET Shanghai.

The construct of NMET Shanghai has evolved over the years with the development of

pedagogical and assessment theories. Attempts to measure examinees' speaking skills date back to the early 1990s under the influence of the Communicative Language Ability (CLA) model (Bachman 1990; Bachman & Palmer, 1996). An optional test of spoken English has been held separately since then. In the first few years, examinees talked to human interlocutors, then the test became computer-based in 2000. The score of the oral test served as a reference for college admission to foreign-languages-related majors. It was in 2017 that a listening and speaking test was made mandatory for all Shanghai examinees.

As a result, NMET Shanghai currently consists of a written test of listening, reading and writing items and an independent listening and speaking test to be administered after the written test. This review focuses on the listening and speaking test only.

Test structure

The computer-based semi-direct listening and speaking test of NMET Shanghai is carried out in standardized language labs with examinees wearing headphones and having their oral production recorded.

The test is divided into two parts: a speaking part in which examinees do not need to listen for extra information and an integrated listening and speaking part where examinees produce a response based on what they have heard. There are six sections, as shown in Table 1.

Table 1. Content and format of the listening and speaking test

Part	Task format	Pattern of interaction	No. of items	Preparation/ response time
Speaking	Section A. Read aloud sentences	Monologic	2	60/30 seconds
	Section B. Read aloud a passage	Monologic	1	60/30 seconds
	Section C. Ask questions	Interactive	2	0/40 seconds
	Section D. Describe a four-panel comic	Monologic	1	60/60 seconds
Listening and Speaking	Section A. Make a quick response	Interactive	4	0/20 seconds
	Section B. Listen to a passage and answer questions	Monologic	2	30/30 seconds for Q1, 60/60 seconds for Q2

In the speaking part, Sections A and B measure examinees' mastery of pronunciation. In Section C, examinees ask two questions about a given situation to obtain necessary information (e.g., "Your classmate has got a football with some football stars' signatures. Ask him about those signatures."). At least one *wh*-question should be raised (e.g., "Where have you met so many football stars?"). This task aims at measuring examinees' communicative competence in given contexts (Liu & Chen, 2018; Xu, 2021), though no response will be given to examinees' questions. Section D tests the ability to develop a few coherent sentences.

The listening and speaking part is also made up of independent and interactive tasks. Section A consists of four questions or statements. Examinees are required to produce a proper response to each of them (e.g., on hearing "Linda won the championship again in the tennis game!", a preferable response is "Congratulations!"). Then in Section B, examinees answer two questions after hearing a 200-word passage that is played to them twice.

As shown in Table 1, test takers are given time to prepare for monologic tasks but not for interactive tasks, so as to mimic the quick responses expected of real-life conversation. Between each section, there is a short break of 10 seconds. The whole

test lasts about 20 minutes.

Test administration, scoring, and score reporting

A commissioned office (officially renamed Shanghai Municipal Educational Examinations Authority, SMEEA, in 1995) oversees the development and administration of NMET Shanghai.

The test takes place in standardized labs, where examinees wear headphones and have access to a display. Throughout the test, instructions (in English) are read to the examinees and appear simultaneously on the screen. To make "talking to a computer" less awkward, an image of a real person appears on the screen as an interlocutor to give instructions and speak to examinees.

Currently, the listening and speaking test accounts for 10 points out of the total 150 points in NMET Shanghai. The test is scored by both humans and AI-machines (in-house scoring engines) (Xu, 2019). If the human-rater and the scoring engine assign two scores with too big a gap, a second human-rater decides the final score. However, the score of this test is not reported to examinees separately, who only receive a total score for the written test and listening and speaking test combined.

Appraisal of the test

Test construct

The listening and speaking test makes use of six different item formats to measure examinees' pronunciation, intonation and stress, ability to sustain a monologue, and more importantly, communicative competence. However, the current test design does not work well for this purpose.

First, four out of six tasks of the current listening and speaking test are monologic, and they take up too much of the test. Although scores from monologic tasks may predict

scores on tasks designed to measure learners' interactional abilities to a moderate extent, their predictive strength is lowest for test takers close to university entry (Roever & Ikeda, 2022).

Another problem with the current test lies in its unnatural way of simulating interactions. There are merely two interactive tasks in the test: asking questions about a given situation (Speaking Section C) and giving response to a particular sentence/statement (Listening and Speaking Section A), which are intended to measure sociolinguistic competence and general language ability. Specifically, the "asking questions" task has examinees propose questions, similar to what an interviewer would do. In the "giving responses" task, however, examinees become the ones responding. Such a design deliberately splits a real-life-like conversation into two separate tasks, while spoken interaction is supposed to be dynamic and co-constructed and shared between interlocutors (Galaczi & Taylor, 2018). Turn-taking, which is common and natural in daily communication, is also made impossible. However, in defense of SMEEA, even the speaking test of the Test of English as a Foreign Language internet-based test (TOEFL iBT) is essentially non-interactive (Johnson, 2001; Roever & Kasper, 2018), suggesting that the underrepresentation of interactional competence in NMET Shanghai's listening and speaking test is perhaps not unique. Unfortunately, a ready-made solution is unlikely to be available in the near future.

SMEEA has acknowledged that examinees cannot yet converse naturally with a computer and reported that it was finding ways to engage candidates in a more authentic multi-turn conversation with possibly an AI interlocutor (Xu, 2021, p. 26). With the rapid development of artificial intelligence, a practical solution may eventually be found to make the test more authentic.

Furthermore, there are other aspects of speaking that cannot be measured by merely having examinees taking part in a conversation, such as how to manage discourse, how to make meaningful negotiation, how to take turns as naturally as possible, etc.

Therefore, the Zheng and Xu (2019) claim that "the construct of NMET Shanghai was complete with the incorporation of the listening and speaking test" (pp. 34-35) is arguable.

Test reliability

When the listening and speaking test became mandatory, the large number of test takers necessitated multiple parallel tests. According to Xu (2021), 10 parallel listening and speaking tests were used in the NMET Shanghai test in July, 2020. As Chinese students "invest considerable time and effort in assessment-related preparation activities, such as memorization of material and practicing on past test papers" (Carless & Lam, 2014, p. 170), the recycling of used test items is impossible. The challenge, therefore, is to balance the difficulty level of all parallel tests.

According to SMEEA, three tactics have been adopted to meet this challenge. First, a team of test developers with advanced degrees in relevant fields and with extensive experience in language learning and assessment have been working closely together to ensure the quality of the test. Second, very detailed test specifications were made before any parallel test tasks are created. Take the "reading aloud" task as an example: sentences to be read by examinees should be of very similar length, contain the same number of words and syllables, and be expressed with similar intonations in all parallel tests. Third, in one test paper, topics should be diversified; while in parallel tests, all topics should be general enough and familiar to the examinees. This helps to enhance test fairness and avoid bias. Then comes the problem: with 12 items across six tasks, covering 11 topics (the last two items are based on the same text), it is not likely that all parallel tests can cover the same topics, making it a game of luck, which is also true of many existing tests.

However, an even more major problem with test reliability lies in it being untransparent. In fact, almost no research studies/technical reports have ever been published concerning the development of NMET Shanghai or its sub-tests. Although

much data related to the test has been carefully collected over the years, it remains unavailable to independent external researchers. Even the validity of the test could be doubted. In an age of collaboration and test accountability, it is probably time to make a change.

Test washback

After listening comprehension became part of the written test in 2001, senior secondary school English teachers and administrators began to allot considerable time to coaching students in listening comprehension (Xu, 2005). It was hoped that the incorporation of the speaking tasks in 2017 could bring similar washback to the learning and teaching of English speaking skills.

Schools have allocated more time to the teaching of speaking since 2017. Hou (2018) surveyed 327 *gaozhong* (senior secondary school) English teachers and found that almost half of the schools involved had done so. Xu (2021) found that some senior secondary schools in Shanghai continued to offer English speaking classes for senior year students, even when they were overloaded due to test preparation.

Studies have found that *Gaozhong* teachers welcomed the change and considered it necessary, and as a result, most of the teachers surveyed spent more time on the training of listening and speaking skills than before (Liu & Chen, 2018; Zhang, 2019). Nonetheless, most of the training offered by *gaozhong* teachers was test oriented. Having students work on sample tests and reading textbooks aloud were teachers' favorite class activities (Cheng et al., 2021; Zhang, 2019), yet they were unhelpful for cultivating students' speaking competence.

On the bright side, *gaozhong* students were found to be more encouraged than intimidated by the test. Xu (2021) found that as learners, students had become more aware of the importance of improving their speaking skills. Liu and Chen (2018) interviewed 82 students and found that the majority of them had invested one more

hour into speaking activities every week and were more confident with their speaking performance than they were before the launch of the test. Cheng et al. (2021) found that students from first-tier schools (also known as key senior secondary schools in Mainland China) were more willing to participate in activities like giving speeches and presentations.

Overall, these studies show that the implementation of the test did motivate examinees to practice spoken English when they were preparing for the test, though it is not yet clear exactly how effective this was in improving their oral English proficiency. There is still a paucity of washback research, and further study on the impact of the test is still urgently needed.

Summary

Though not perfect, the listening and speaking subtest of NMET Shanghai has been generally positive in promoting the learning and teaching of communicative language ability (Zheng & Xu, 2019). As it is one of China's first speaking and listening tests, its development and administration could lead to more effective speaking assessments in the near future. As a firm attempt to fight "dumb English" in the country, the impact of the test also brought to other provinces a wealth of experience and insight. To accommodate new generations of examinees, more effort will need to be put into upgrading test tasks for better construct coverage of communicative competence, speeding up the building of an even larger pool of parallel test items for the large number of test takers, and drawing more researchers to validate the test and study its washback.

Reviewed by *Mingwei Pan,* Shanghai International Studies University *& Yang Wang,* Shanghai International Studies University / Shanghai Maritime University

Acknowledgements

This work was supported by the research project "A multi-perspective validation study on tests for English majors" of Shanghai International Studies University [2022113034].

References

Bachman, F. L. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Bachman, F. L., & Palmer, A. (1996). *Language Testing in Practice*. Oxford University Press.

Carless, D., & Lam, R. (2014). Developing assessment for productive learning in Confucian-influenced settings: Potentials and challenges. In Wyatt-Smith, C., Klenowski, V., & Colbert, P. (Eds.) Designing Assessment for Quality Learning (pp. 167-179). Springer. https://doi.org/10.1007/978-94-007-5902-2_11

Cheng, X., Zhang, S., & Qian, J. (2021). The washback of the listening and speaking component in the new English Gaokao in Shanghai. *Foreign Language Learning Theory and Practice* (03), pp. 83-94.

Galaczi, E., & Taylor, L. (2018). Interactional Competence: Conceptualisations, Operationalisations, and Outstanding Questions. *Language Assessment Quarterly* (15), pp. 219-236. http://doi.org/10.1080/15434303.2018.1453816

Hou, Y. (2018). A study on the washback effect of the reform of SHMET Listening and Speaking Test. *Technology Enhanced Foreign Language Education* (183), pp. 23-29.

Johnson, M. (2001). *The art of non-conversation. A reexamination of the validity of the oral proficiency interview*. Yale University Press.

Liu, S., & Chen, Y. (2018). A practical exploration on NMET (Shanghai)-based English listening and speaking teaching. *Technology Enhanced Foreign*

Language Education (183), pp. 30-34.

Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*, 39(1), pp. 7–29. https://doi.org/10.1177/02655322211003332

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, 35(3), pp. 331–355. https://doi.org/10.1177/0265532218758128

Xu, W. (2019). Achieving test fairness: In the College Entrance Examination (Shanghai English Paper). *Foreign Language Testing and Teaching* (02), pp. 9-16.

Xu, W. (2021). Practice of speaking assessment in large-scale high-stake examination: A case study of Shanghai English Gaokao Listening and Speaking Test. *Foreign Language Testing and Teaching* (01), pp. 21-27.

Xu, X. (2005). A practical study on computer-assisted English Oral Test of University Entrance Exam in Shanghai. *Shanghai Research on Education* (10), pp. 52-55.

Zhang, R. (2019). Backwash effect of integrating listening and speaking test into NMET (Shanghai): Taking School J as an example. *Foreign Language Testing and Teaching* (04), pp. 47-53.

Zheng, F., & Xu, W. (2019). Advancing English language education in Shanghai with the new College Entrance Examination. *China Examinations* (9), pp. 32-36. https://doi.org/10.19360/j.cnki.11-3303/g4.2019.09.004