

Article

Is the OPI Useful? A Comparison of L2 Chinese Learners' Performance and Perceptions of ACTFL OPI and HSKK Oral Proficiency Tests

Jun Wang

Shanghai Jiao Tong University, China

Clare Wright*

University of Leeds, Leeds, UK

Na An

Shanghai Jiao Tong University, China

Sicheng Wang

Macau University of Science and Technology, China

Received: 4 December, 2024/Accepted: 28 February, 2025/Published: 17 March, 2025

Abstract

The ACTFL oral proficiency interview (OPI) test is widely adopted for second language (L2) assessment, with claimed benefits of eliciting fluent, realistic dialogic speech, and fostering a motivating learning experience. However, researchers have not reached a consensus on operationalizing its usefulness (Bachman & Palmer, 1996), particularly in ways that would enable direct comparison between test-taker performance and perceptions of the ACTFL OPI and in non-interview type L2 oral tests (e.g. IELTS for English or HSKK for Chinese), leaving a gap which needs addressing. Among non-OPI assessments, English tests such as IELTS have been widely evaluated, but HSKK remains under-investigated, despite the rapidly growing number of L2 Chinese learners, suggesting that closer investigation is needed of how different formats may impact on L2 Chinese test-takers' oral proficiency and their attitudes towards the tests. The current study aims to fill these gaps in a novel mixed-methods case-study, investigating comparability between the ACTFL OPI and HSKK, in relation to range of outcomes in test-takers' performance, to proficient and authentic use of linguistic knowledge, and perceptions of test experience. Participants were forty Chinese second language learners of various first language backgrounds and proficiency levels, previously unfamiliar with either type of test. Each participant was tested using standard ACTFL OPI and adapted HSKK procedures by professional testers over a 2-day testing period. Speech data were analysed to assess a) dispersion across ratings, b) proficiency in performance, using both objective indicators of complexity, accuracy and fluency, and evidence of "real-life" authentic speech. In addition, a survey and a group interview were conducted to gather data on participants' experiences and perceptions of taking the tests. Results confirmed that the ACTFL OPI ratings were not as widely dispersed across test-takers as non-OPI scores; the ACTFL OPI speech samples elicited more variety of linguistic

*Corresponding author. Email: c.e.m.wright@leeds.ac.uk

proficiency, and more real-life authentic speech style; the ACTFL OPI was also deemed more positive with potential washback effects toward more motivated learning experiences. This study reveals both limitations and advantages in using ACTFL OPI compared to non-OPI tests, particularly for L2 Chinese; overall the findings echo current pedagogic moves towards encouraging more authentic communicative speech abilities, both for assessment and real-life use.

Keywords

ACTFL OPI, oral test effectiveness, Chinese, HSKK

1 Background

The Oral Proficiency Interview (OPI) of the American Council on the Teaching of Foreign Languages (ACTFL) and its accompanying Language Proficiency Guidelines (ACTFL, 2012, 2018) has developed into a globally recognised high-stakes proficiency test for many of the world's languages, with both firm proponents (e.g. Liskin-Gasparro, 2003) and critics (e.g. Bachman & Palmer, 1996; Salaberry, 2000). Briefly, the OPI measures oral proficiency through a structured interview lasting around 15-30 minutes, either face-to-face or online (Isbell & Winke, 2019). The interview is conducted by certified testers, following established protocols, who assess the speaker's ability to perform language tasks across various contexts and content areas. The OPI is structured into four phases: a warm-up phase to set the interviewee at ease, checks to establish general proficiency level, probes to test the range of the speaker's linguistic knowledge and speaking ability, and a wind-down stage to end the exam on a positive note. During the check and probe stages, the tester elicits language samples which are then rated according to the ACTFL Proficiency Guidelines, as Novice, Intermediate, Advanced, and Superior, with further sub-levels within each category. Protocols around identifying different proficiency levels aim to ensure as much objectivity as possible, but to some extent a tester has to rely on manipulating these two phases and using personal judgment to identify the most proficient speech samples, so as to assign a sub-level out of the 10 official levels as regulated by the ACTFL Proficiency Guidelines. The test is widely used in academic, governmental, and professional settings to assess language proficiency for purposes such as college class placement, certification, and employment (ACTFL, 2012).

In the forty years since its initial publication in 1982, scholars have debated extensively about the test's usefulness and general validity without reaching a final consensus (e.g. Bachman & Palmer, 1996; Carey et al., 2011; Chalhoub-Deville & Fulcher, 2003; Halleck, 1992; Henning, 1992; Isbell & Winke, 2019; Johnson, 2000; Omaggio, 1986; Surface & Dierdorff, 2003).

The OPI, according to its supporters, has been improved since its earliest inception and now provides more consistent positive advantages in response to the challenges identified by its original detractors. First, the OPI format offers high face validity due to its flexible, communicative, testee-focused nature. As indicated in its official tester training manual, the test "is an interactive, adaptive and speaker-centered assessment"; "the OPI adapts to the speaker's level of proficiency by limiting the range of linguistic tasks to those that the speaker can manage more or less successfully. The topics treated in the OPI are based on the interests and experience of the speaker" (Swender & Vicars, 2012). Its official examinee handbook (ACTFL, 2018) claims that it elicits a speech sample that assesses examinees' ability to communicate in the target language (proficiency) while performing the functions that one might encounter in real-life situations. The "adaptive", "speaker-centered" and "real life-like" characteristics make it distinctive from other standardised oral proficiency tests such as the English IELTS test or, for Mandarin, the Hanyu Shuiping Kouyu Kaoshi (HSKK). These claimed features offer several notable benefits: 1) "every testee would be tested with materials and tasks that best fit his/her proficiency level", and the test result might better disperse all participants, which is of significant importance if the test is used for high-stake

purposes; 2) the test might result in participants producing more natural and real-life like speech, which fit the its claimed purpose; 3) the test could have positive washback effects by encouraging the wider second language education community to move away from performative exam-oriented tests towards learning language skills that are effective in everyday use (Isbell & Winke, 2019). However, while the ACTFL OPI has clearly gone some way to address original critiques, there remain two principal areas of debate that drive this current study.

2 The Necessity of a New Comparative Study

First, we lack clear empirically-grounded agreement on how to operationalize its claimed linguistic benefits for speech performance, particularly in ways that would enable direct comparison of oral proficiency between the ACTFL OPI, on the one hand, and other official oral proficiency tests (usually provided and endorsed by the countries that use the test languages as their L1) in non-OPI form – for Chinese, this is the Hanyu Shuiping Kouyu Kaoshi (HSKK). Therefore, linguistically-informed studies comparing the ACTFL OPI and its official non-OPI counterpart would help to validate the former’s aforementioned suitability for developing oral proficiency. While the problem is not so great for English tests such as IELTS, which have been widely evaluated, tests in other languages such as Mandarin remains under-investigated, despite the rapidly growing number of L2 Chinese learners (Zhang & Lin, 2017).

Two further elements require further investigation to ensure the comparisons will be appropriate. First, OPI tasks need sufficient difficulty to fully disperse a heterogenous group of test takers (Fulcher, 2014; Stansfield and Kenyon, 1992), so OPI must be evaluated whether it is a suitably difficult task by comparison with non-OPI tasks. Second, effectiveness in eliciting and rating “real life-like speech” is hard to validate, since the kind of speech itself is not easy to identify. If we compare more “real life-like” speech samples with more “exam-like” tasks (e.g. describing pictures, answering questions after preparation and retelling stories), two distinctive aspects of “real-life” speech can be seen as being unprepared and dialogic. These features can be further analysed in more detail using objective linguistic frameworks, such as the CAF (complexity, accuracy and fluency) Framework (see Housen & Kuiken, 2009). Several studies (e.g. Skehan and Foster, 1999; Wright, 2020) have found that speech samples elicited under unprepared conditions were less complex. Michel, Kuiken and Vedder (2007) found dialogic tasks triggered more accurate and fluent output, though it was structurally less complex. Tavakoli (2016) also confirmed that dialogues can foster better fluency, though this can depend on type of dialogue (Wright, 2020). We might hence assume that if the ACTFL OPI does elicit truly dialogic and unprepared speech samples, comparatively higher fluency, greater accuracy and lower complexity should be detected compared to non-OPI tests, sustained across both objective CAF measures and rubric-based OPI-style holistic ratings. These issues and gaps prompt our first three research hypotheses.

Second, debates also remain over learner perceptions of different test formats, and potential washback effects (Malone & Montee, 2010; Thompson et al., 2016). Such debates have practical and financial implications for institutions in terms of test selection and tester training, as well as ensuring suitable student accessibility and positive learning experiences. Washback effects can be seen in Chinese contexts, where learners who focus on test preparation may demonstrate effective rehearsed performative competence in standardized prepared exam formats, but lack more spontaneous creative competence for unprepared interactions (Wright, 2020), and can struggle to maintain motivation or confidence in real-life conversations (Wright et al., 2022; Zhang & Du, 2023). Learner perceptions of such washback effects and impact on test preparation or learning experiences are important to investigate, particularly in a language like Mandarin which is experiencing very rapid growth in learners worldwide (currently estimated to be around 25 million). To date, there has been little consistent empirical support for claims of the OPI’s positive washback effect on teaching and learning (Isbell & Winke, 2019;

Omaggio, 1986), so it is important to include evidence from test takers' perspectives who have taken both types of oral tests, including if they perceive OPI as a test which encourages more real life-like speech and induces more positive learning attitudes. This gap drives the final hypothesis for our study.

3 The Study

3.1 Four hypotheses

If the ACTFL OPI can live up to its claimed advantages as indicated above, the following hypotheses should be validated:

- Hypothesis 1:** The ratings of the ACTFL OPI are more dispersed than those of non-OPI oral tests for the same group of examinees, if graded according to set OPI rubrics by certified testers, since OPI provides more proficiency-adaptive elicitations.
- Hypothesis 2:** The speech samples elicited in the effective “probe” phases of ACTFL OPI are more dispersed in quality under objective CAF analysis, i.e., greater range across CAF measures, compared to those elicited in non-OPI oral test tasks, since OPI provides more proficiency-adaptive elicitations.
- Hypothesis 3:** The speech samples elicited in ACTFL OPI are dialogic and unprepared, so will be less complex, but more accurate and fluent, in comparison to those elicited in non-OPI oral test tasks. These characteristics emerge because the ACTFL OPI elicits authentic “real life-like” and communicative speeches.
- Hypothesis 4:** The ACTFL OPI is perceived as reflecting more real life-like language use and will be better accepted than non-OPI oral tests regarding aspects that promote second language learning by examinees who take both types of tests under unbiased conditions, since the former sets positive “washback” effects toward teaching/learning as its aim, and tries to achieve it with its unique design.

3.2 Participants

Invitations to participating the current study were sent to all learners enrolled in a summer program of Chinese as a second language covering all proficiency levels at a Chinese university. For ecological validity and authenticity reflecting typical test-takers in such settings, as well as for practical reasons given the complex nature of the study, we did not set L1 background or proficiency level as a specified variable. Forty learners were selected through random sampling out of all 96 respondents to become the participants of the study. Institutional ethical protocols were followed to ensure informed consent, confidentiality and anonymity, and secure data management. The mean age of participants was 24.9 (SD=4.55), gender distribution was fully balanced at 20/20. The sample comprised Korean (12), English (6), German (5), Spanish (3), Thai (3), Japanese (3), French (2), Russian (1), Filipino (1), Turkish (1), Indonesian (1), Vietnamese (1) and Kyrgyz (1). At the time of study, the mean length of learning Chinese was 18.07 months (SD=16.78 months).

3.3 Test materials

More importantly, none of the participants had ever taken any of the two tests, nor had even heard of them, ensuring that there would not be confounding test practice effects. Materials were designed on the basis of the two tests. We use **OPI** and **Non-OPI** to stand for the two types of oral test tasks being compared here, although OPI was slightly modified in form to serve the purpose of the study. The term Non-OPI was used instead of HSKK since there are significant modifications from its original form. The reasons for modifications will be elaborated below.

3.3.1 OPI testing

One ACTFL-certified examiner was invited to administer a standard complete face-to-face OPI test to the participants, with all phases and features such as role play cards preserved. One hidden change was added to the procedure: the examiner always said the same code phrase “xia yi ge wen ti” (meaning “next question” in Chinese) once she began to “probe”. All tests were recorded with a computer and saved as audio files for analysis. After completing all testing, two certified testers, including the one who administered the testing, served as the analysts of the OPI data. Because of the existence of the code phrase and the full participation in testing, they can easily identify the four stages of OPI (including the ending of the “probe”, since it was determined by the examiner) as well as the role plays administered in the test. Those stages and role plays were then truncated and saved as separate audio files (each stage with multiple files if it happened more than once) for further analysis.

3.3.2 Non-OPI testing

This part was modified from the original form of HSKK to serve the purpose of the current study. To explain the modification, a brief introduction to HSKK is presented here. The test has three main levels, namely primary, intermediate and advanced, which provide their respective tests. An examinee will have to decide which level to take at registration. The test results fall into “fail”, “pass” or “excellent” for each level. HSKK is tape-mediated and include six types of tasks across all levels: T1: repeating the sentence after listening; T2: answering a (short) question instantly after hearing it; T3: retelling a story after listening to it; T4: Describing a picture after preparation time; T5: answering (long) questions written on the paper after preparation time; T6: oral reading a long passage or text. Primary level delivers T1, T2 and T5; intermediate level delivers T1, T4 and T5, while the advanced level delivers T3, T6 and T5.

Participants in real-world HSKK exam contexts register for one specific level. In our study, aiming for equivalent level-probing in both contexts, this was not possible. One simple solution to test an examinee of any possible level would be to deliver all three complete tests to him/her and place him/her at the highest level that his/her performance can hold. However, this would make the testing time more than 90 minutes in length, since it would include 9 tasks, 47 items plus a lot of preparation time between tasks. To solve this problem, in the study, we collapsed the overlapping tasks and items across the three levels, and reached a test containing 6 tasks (T1 to T6) and 28 items. Different difficulty levels within the same task for T2 and T5 were represented by corresponding items (see Appendix 2). After the modification, the test became 40 minutes in length including preparation time, which is only slightly longer than a typical OPI test. Using the official grading criteria issued by Chinese Testing International (CTI, 2020), the outcome of the test could be assessed after re-splitting it into three parts: Primary level (5 primary items in T1, T2 and 1 primary item in T5), Intermediate level (5 intermediate items in T2, T4 and 1 intermediate item in T5) and Advanced level (T3, 1 advanced item in T5 and T6). The outcome in the highest passed level (pass/excellent) became the HSKK grade for the examinee, overriding the outcomes in other levels. This removed the need for the “fail” category other than at the first primary level.

A potential concern over collapsing tests to reduce time is that this could have reduced the validity of the grading. Our three reasons to justify this decision are as follows. First, a 90-minute test would possibly make examinees generate negative feelings against the test, which would not only affect the reliability of the language samples elicited, but also interfere with the results regarding hypothesis 4. Reduction in length thus became necessary. Second, all item types in the original HSKK test were kept, while the results retrieved from three main levels can cross-validate each other to ensure accurate grading. This is similar to the way of reaching a grade in the SOPI tests as mentioned in part 2. Third, we believe the current task design and official grading system of HSKK is not perfect. Task filtering and simpler grading as introduced here could, we argue, improve the quality of testing process (see part 3.5

below). The 6 tasks in the non-OPI part of the study all demonstrate typical non-OPI features, which are tape-mediated, monologic, fixed in form and content, and mostly prepared. In our study, the test materials were fed into a computer program that could play the directions back to test takers and record their oral productions, along the lines of a OPIc. Waiting/preparation time required by the test was also embedded into the program to maximally mimic the situation of a real HSKK test. We believe therefore that our minor changes to the HSKK testing format were appropriate to increase comparability of the two types of testing at lowest cost to methodological validity.

After all the recordings were collected, a research assistant collated them into 6 audio files representing T1-T6 respectively for analysis.

3.3.3 Survey

To collect participants' data of demographic information and data addressing hypothesis 4, as well as self-report data related to hypothesis 1-3 that could provide triangulation, a survey (see Appendix 3) was administered to all participants after they took both oral tests (see part 2.4 for procedures). Besides background demographic data (omitted here), the main body of the survey consisted of 8 pairs of 5-degree Likert scale questions investigating participants' perceptions of various aspects of these two types of tests. There were also two open questions asking participants to report perceived advantages and disadvantages of the two tests.

3.3.4 Linguistic resource test

To better account for the possible findings, participants' vocabulary size and syntactic knowledge were also tested after the survey. For vocabulary size testing, stratified sampling was used to select 30 words from the 6000-word official HSK vocabulary list. The accompanying multiple-choice questions were meaning-translation tasks. Participants' scores ranged from 0 to 30, which can be used to estimate their vocabulary size if multiplied by 200. The syntactic knowledge test included 20 multiple-choice questions covering the major grammar points of Mandarin Chinese to generate a syntactic knowledge score ranged from 0 to 20. Shi's (1998) study on acquisition of common syntactic structures was referred to when this section was designed.

3.3.5 Group interview

After all participants completed both oral tests, the survey and the linguistic resource test, 4 of them were chosen through random sampling to participate in a group interview, which is guided by the 5 questions as listed in Appendix 4. This part was another means to validate hypothesis 4 and to explain the findings of the study. The interview was about 35 minutes in length, and was recorded and transcribed to text for analysis.

3.4 Procedures of data collection

The data collection went on for two weeks, however, each participant's data were collected in two consecutive days after he/she was scheduled by the research assistant. All 40 participants were evenly split into group A and B through random sampling. They were briefed with the general formats and tasks of each test by a research assistant right before they took it. Group A took the OPI test on day 1, and the non-OPI test on day 2. They also filled out the survey and the linguistic resource test on day 2, after completing the oral test. Group B went through parallel procedures, but with the order of OPI and Non-OPI reversed. A group interview with the 4 chosen participants was administered by a research assistant on the last day of the two-week period.

3.5 Data analysis methods

3.5.1 Grading the tests

Four bilingual Mandarin-English raters were recruited to assess the data; two were officially ACTFL-certified raters to assess the OPI data, and included the test examiner, as noted above; the other two were HSKK-certified to evaluate the HSKK data. They were also employed as language teachers at the host university, but were not known to the participants.

The complete recordings of the OPI tests were graded separately by the two ACTFL-certified examiners. They each placed a participant at a level ranged from novice low to superior following standard grading procedures and rubrics as regulated by the official tester training manual. The 10 levels were also converted to a numeric scale of 1 to 10 as shown in Table 1 below. Mean value of the two scores was used as the finalized OPI score of each participant.

Table 1

OPI Levels Converted to Numeric Scores

OPI level	NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Score	1	2	3	4	5	6	7	8	9	10

Key: NL=novice low, NM=novice mid, NH=novice high, IL=intermediate low, IM=intermediate mid, IH=intermediate high, AL=advanced low, AM=advanced mid, AH=advanced high, S=superior

Two different methods were employed to rate the non-OPI recordings. First, the official HSKK grading rubric was referred to by the two certified HSKK testers who graded the recordings separately. The HSKK grading rubrics gave “pass” or “excellent” to each of the three primary levels, with an additional “fail” rating for those who failed to reach the primary pass level. The ratings in each main level were then converted to a finalized HSKK score within a continuous scale ranging from 1 to 7 as shown in Table 2. Those who failed the primary level test were given a “1” since “0” should be total silence according to the official rubric. In this way, a more manageable combined test could be carried out, instead of three separate tests at different levels; each participant’s score was recognized as valid at the highest main level reached by the participant, while the scores achieved in lower main levels were overridden by the former.

Table 2

Converting HSKK Levels to Numeric Scores

HSKK level	PF	PP	PE	IP	IE	AP	AE
Score	1	2	3	4	5	6	7

Key: PF=primary fail, PP=primary pass, PE=primary excellent, IP=intermediate pass, IE=intermediate excellent, AP=advanced pass, AE=advanced excellent

There were two reasons that another way of grading should be adopted. First, the previous way of grading did not provide a set of universal criteria for grading oral performance; second, T1, T2 and T6 have been widely criticized for lacking validity. Hence our second method for rating the non-OPI part worked holistically and only focused on T3, T4 and T5.

The band descriptors of the IELTS (International English Language Testing System) speaking test (IELTS website, 2016) were borrowed to form a 4-category, 9-scale marking rubric. This choice was made for a few reasons. First, the IELTS speaking was perhaps the most mature and widely accepted non-OPI test; second, its grading rubrics assess (1) fluency and coherence, (2) lexical resources, (3)

grammatical range and accuracy, (4) pronunciation, which were not only “non-English specific”, but also highly related to the objective analysis of language samples (see below). Third, its wider range (9 points compared to 7 points in HSKK) also made it more comparable to OPI (10 points) in regard of the capability of spreading participants. In fact, a similar attempt had been made by Wang et al. (2018) in an oral proficiency related study. Both the HSKK rating (method 1) and IELTS rating (method 2) were used in data analysis below, which can also validate each other.

3.5.2 Objective analysis of the language samples


To test hypothesis 2 and 3, objective analyses of speech samples elicited in both types of testing were needed. Considering the characteristics of Mandarin Chinese and the length/types of the speech samples, the following 9 sub-indices were used to analyze the speech samples. For syntactic and lexical **Complexity**: words per AS-unit (Analysis of Speech Unit, Foster et al., 2000), clauses per AS-unit and type-token ratio (Foster et al., 2000). For **Accuracy**: pronunciation (consonant/ vowel/tone) errors per AS-unit, lexical errors per AS-unit, syntactic errors per AS-unit (Yuan & Ellis, 2003). For **Fluency**: total syllables per minute, mean length of run (number of syllables between pauses), pauses per minute (Tavakoli, 2016).

The following five audio clips were truncated from the original recordings of the two types of testing from each participant to enter comparison. From OPI: **RLC** (Random Level Check): a random round of level check, defined as a complete piece of utterance elicited by a single question or direction from the tester, which was not marked with the code phrase; **BPR** (best probe): the best utterance elicited in the “probe” phase of the OPI, defined as the speech sample of highest proficiency level that a participant was able to maintain (without obvious linguistic breakdown). The breakdown appears in the language samples elicited by a probe at the next higher level, and was elicited by the question or direction marked with the code phrase. According to the OPI’s grading rubrics, the BPRs were actually the major determining elements for level placement. The location and extraction of RLCs and BPRs were done manually by the OPI tester who had administered the test. Since there were usually more than one BPR for each participant at the same highest proficiency level, which cross-validate each other as required by standard testing procedure, the longest one was chosen by the tester for analysis in the study. From non-OPI: although all six tasks had already been extracted from the recording, and coded to retrieve CAF indicators (see below), the data from T1, T2 and T6 were not included in cross-task comparison for the reason indicated above. Among “core tasks”, **RTS** (T3, retelling stories) and **PD** (T4, picture description) both involved the combined data of two items in each task since they were at the same difficulty; **ALQ** (T5, answering long questions) only involved the data from one item at the highest difficulty level that was “passed” by the participant out of the three. Table 3 exemplified OPI elicitations and HSKK tasks that elicited the 5 types of speech samples from a participant rated as Intermediate Mid.

Table 3

*OPI Elicitations and HSKK Tasks Used to Elicit 5 Types of Speech Samples from a Participant Rated as Intermediate Mid in OPI**

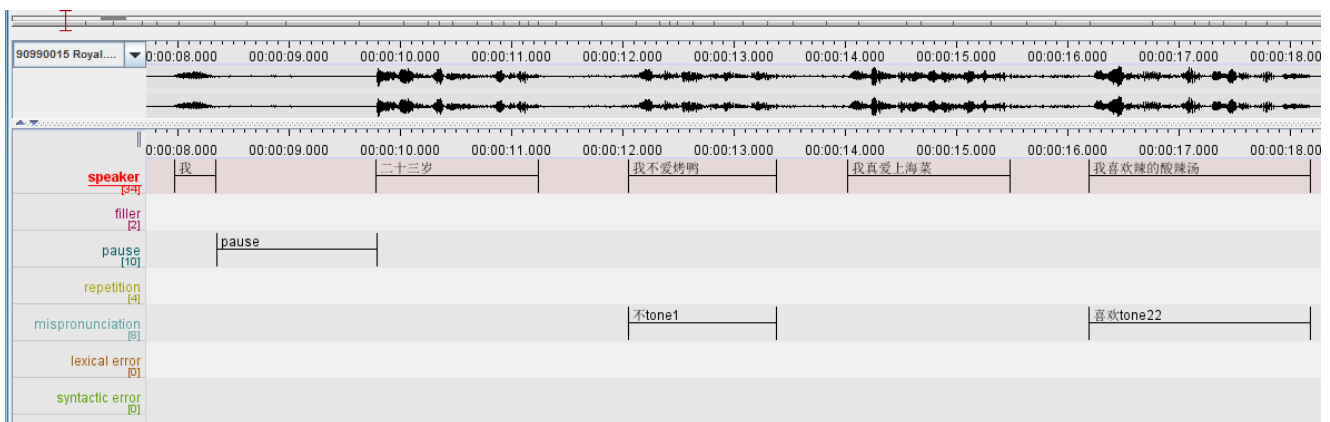
Sample types	OPI elicitations/HSKK tasks
RLC	你觉得看电影有什么好处呢? (What do you think are the benefits of watching movies?)
BPR	那下一个问题是你能把上海和海德堡对比一下吗? (The next question is that could you compare Heidelberg to Shanghai?)

- RTS (Listen and retell the story) 早上，我从衣柜里拿出新买的衣服，对着镜子试了起来。一回头，发现五岁的女儿正在身后看着我，于是我高兴地问她：“妈妈穿这件衣服合适不合适？”女儿从头到脚仔细地看了看，然后说：“衣服很合适，可是妈妈需要再高点儿、再瘦点儿。”
(In the morning, I took newly bought clothes out of the wardrobe, and tried them on in front of the mirror. When I turned back, I found my five-year-old daughter standing there. I asked her happily, “Does the clothes fit mom?” My daughter stared at me from head to feet, and said, “the clothes fit, but mom needs to be a little taller and slimmer.)
- PD 
(describe the picture)
- ALQ 你喜欢和谁一起旅游？为什么？
(Who would you like to travel with? Why?)

**Words in bold are the “code phrase”*

Using the software ELAN, we coded all CAF indicators for each audio clip. The coding was performed by two trained bilingual Mandarin-English research assistants; all coding judgments were cross-checked between the two assistants to reach 100% consensus. Figure 1 below is an example of the coding.

Figure 1
An Example of Objective Coding with ELAN



The CVs (coefficient of variation= (Standard Deviation / Mean) * 100, see Everitt, & Skrondal, 2002) of gradings of OPI and non-OPI were compared to test hypothesis 1. The CVs of all CAF indicators among participants were compared between tasks, with special attention paid to BPR to test hypothesis 2. Repeated measure ANOVAs were performed for all CAF indicators across all 5 tasks to test hypothesis 3.

3.5.3 Analysis of survey, interview and linguistic resource test data

The 8 pairs of Likert scale questions in the survey were put into paired sample t-test to investigate whether there were significant differences in participants' acceptance and preference of the two types of testing. Feedback from the two open questions and the group interview were put into the Nvivo 12 software for qualitative analysis. Results were used to validate hypothesis 4 as well as to provide explanation to the validation/invalidation of hypotheses 1, 2 and 3, so were results retrieved from the linguistic resource test.

4 Results

In regard to hypothesis 1 to check ratings dispersal, we first obtained mean OPI and HSKK scores for each testee. The mean OPI score graded by rater 1 was 4.83 (SD=1.32), that graded by rater 2 was 4.80 (SD=1.16). The frequencies of all sub-levels identified by the two raters were displayed in Table 4. The Cronbach's Alpha of the two OPI raters was 0.952, paired sample t-test showed no significant difference ($t=.298$ $p=.767$), therefore it was reliable to use the mean scores of two raters as the OPI grades for each participant. A histogram was plotted in figure 2 to demonstrate the distribution of the OPI scores.

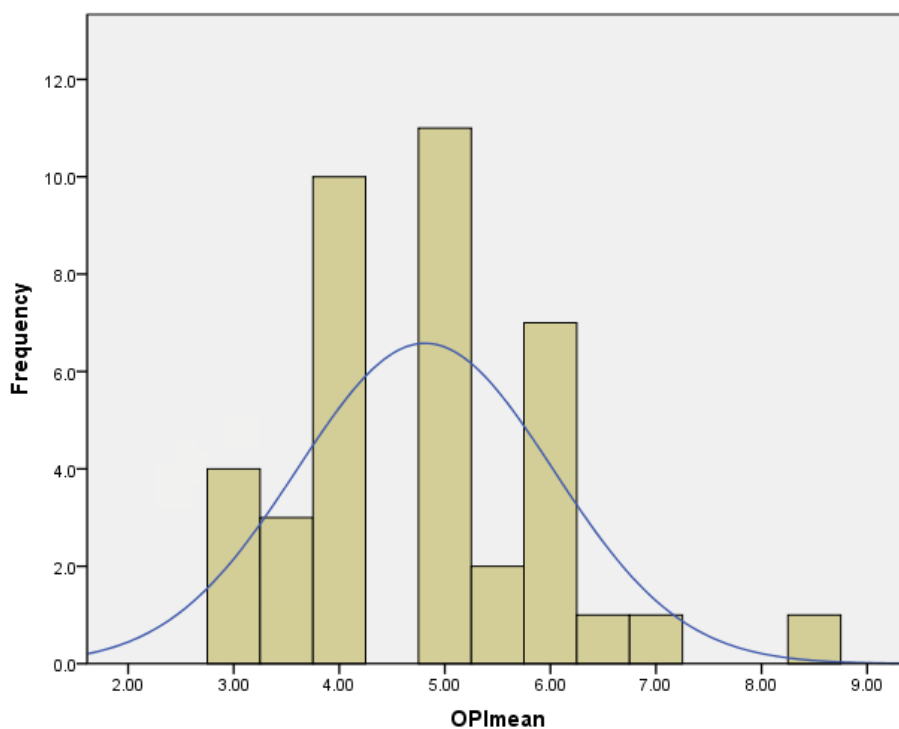
Table 4

Frequencies of All OPI Sub-levels Identified by Two Raters (n=40)

Sub-levels	Numeric score	Rater 1	Rater 2
NL	1	0	0
NM	2	0	0
NH	3	7	4
IL	4	10	13
IM	5	11	13
IH	6	9	9
AL	7	1	0
AM	8	2	0
AH	9	0	1
S	10	0	0

Figure 2

Distribution of Mean OPI Scores



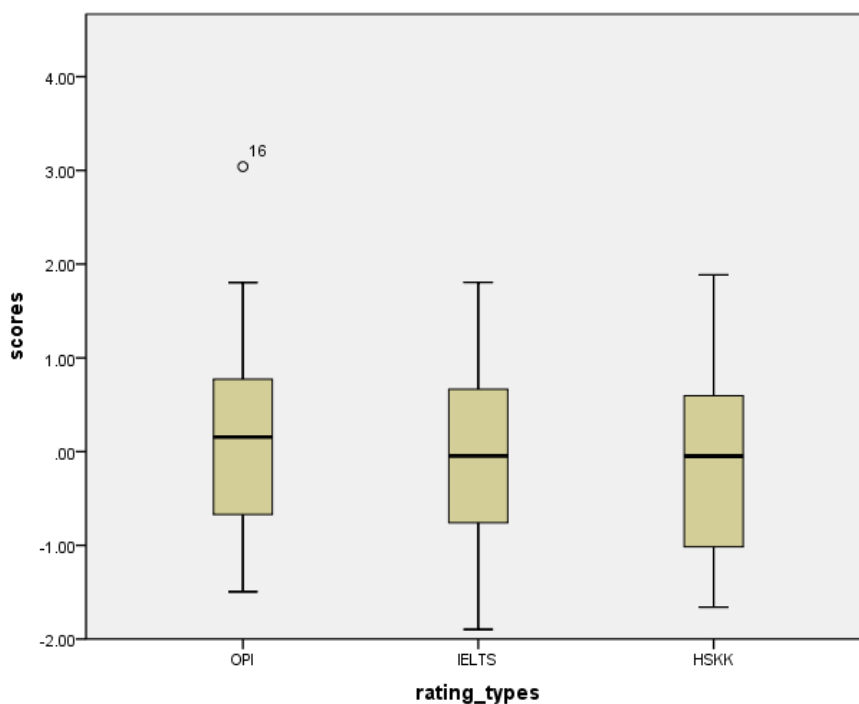
The Cronbach's Alpha of the two HSKK raters was 0.921, and that of the two IELTS raters were 0.950, therefore it was reliable to use the mean scores of two raters as the grades generated by the three types of rating for each participant. A comparison across the three types of ratings were then conducted to check hypothesis 1. However, the results were not quite as predicted. As shown in Table 5, the CV of OPI was 0.25, much lower than that of HSKK (0.43) and IELTS (0.36). To make the results more comparable, we converted all three types of rating into z-scores. The boxplots in Figure 3 showed that the OPI was more ineffective in dispersing participants' oral proficiency compared to either type of rating for the non-OPI test. Hypothesis 1 could therefore not be fully supported, as will be explored in the discussion section below.

Table 5

Dispersion of Three Ratings (n=40)

	Mean	SD	CV
OPI	4.81	1.21	0.25
HSKK	3.58	1.55	0.43
IELTS	4.83	1.76	0.36

Figure 3

Comparison of Dispersion of the Three Ratings Converted to Z-Scores

In regard to hypothesis 2, to check range of speech performance between both tests using objective CAF analyses, we inspected the means, SDs and CVs of all CAF indices, first by examining results combined across the tests, shown in Table 6. The BPR task yielded greatest CV value at 7 out of 9 indices. All complexity and fluency indices were most widely spread out by the effective probe phases in OPI. The case for accuracy was a little complicated. BPR had a very slight advantage in differentiating phonological accuracy (CV=1.01), while lexical accuracy was best dispersed by both RLC and ALQ (CV=1.06), and syntactic accuracy by PD (CV=1.33). These findings, in general, thus support hypothesis 2 that OPI can yield greater dispersal in results than non-OPI tasks when using objective CAF measures.

Table 6
Dispersion of CAF Indices Values across All Tasks

Main category	sub-category	value	Tasks				
			RLC	BPR	RTS	PD	ALQ
Complexity	words per AS-unit	Mean	14.78	19.30	16.79	16.09	22.06
		SD	6.56	8.93	5.47	5.81	6.34
		CV	0.44	0.46	0.33	0.36	0.29
	clauses per AS-unit	Mean	2.43	2.80	2.56	2.44	3.13
		SD	0.78	0.96	0.85	0.63	0.77
		CV	0.32	0.34	0.33	0.26	0.25
	type-toke ration	Mean	0.44	0.51	0.61	0.52	0.42
		SD	0.10	0.13	0.09	0.10	0.09
		CV	0.22	0.25	0.16	0.19	0.22
Accuracy	pronunciation errors per AS-unit	Mean	0.36	0.55	1.61	1.40	1.70
		SD	0.36	0.56	1.20	1.05	1.13
		CV	1.00	1.01	0.75	0.75	0.66
	lexical errors per AS-unit	Mean	0.14	0.31	0.34	0.30	0.27
		SD	0.15	0.27	0.35	0.29	0.29
		CV	1.06	0.86	1.05	0.94	1.06
	syntactic errors per AS-unit	Mean	0.10	0.20	0.16	0.14	0.16
		SD	0.11	0.21	0.19	0.19	0.13
		CV	1.17	1.07	1.23	1.33	0.82
Fluency	total syllables per minute	Mean	128.52	123.88	129.36	112.61	121.31
		SD	44.19	43.42	42.50	35.82	41.89
		CV	0.34	0.35	0.33	0.32	0.34
	mean length of run	Mean	13.18	11.45	8.46	8.94	10.49
		SD	13.11	13.30	9.17	9.18	11.36
		CV	0.99	1.16	1.08	1.03	1.08
	pauses per min	Mean	15.35	14.37	17.13	18.84	16.95
		SD	6.94	7.81	7.34	8.66	7.78
		CV	0.45	0.54	0.42	0.46	0.46

Key: RLC=random level check, BPR=best probe, RTS=retelling stories, PD=picture description, ALQ=answering long questions

Next, for hypothesis 3 to investigate predictions of less complex, but more accurate and fluent speech in OPI tasks compared to non-OPI tasks, one-way repeated measures ANOVAs were conducted to compare

the effects of testing tasks on CAF indices in RLC, BPR, RTS, PD and ALQ conditions, yielding rich details of how each test impacted on speech performance, in ways that analysis of dispersal of general proficiency category ratings could not do. Under the complexity category, significant effects of testing tasks were found on all three indices. For words per AS-unit, Wilk's Lambda=0.40, $F(4, 34) = 12.58$, $p = .000$; for clauses per AS-unit, Wilk's Lambda=0.55, $F(4, 34) = 6.83$, $p = .000$; for type-token ratio, Wilk's Lambda=0.15, $F(4, 34) = 50.31$, $p = .000$. Results of post-hoc comparisons are displayed in Table 7; Figure 4 shows the plotted difference of means of all three complexity indices across the 5 tasks. Across the 9 areas of comparison against non-OPI tasks, RLC was significantly lower (less complex) on 4 of these. BPR was significantly higher (more complex) in 2 groups, insignificantly higher in 4 groups, and was lower in 3 groups. BPR also elicited significantly more complex speech samples in all 3 comparison groups against RLC. Overall, RLC induced lower complexity values in 8 out of 9 comparison groups than non-OPI tasks, and evidence in BPR was very mixed; the OPI thus seemed to elicit less complex speech samples except in its effective probe phases, compared to non-OPI tasks.

Table 7

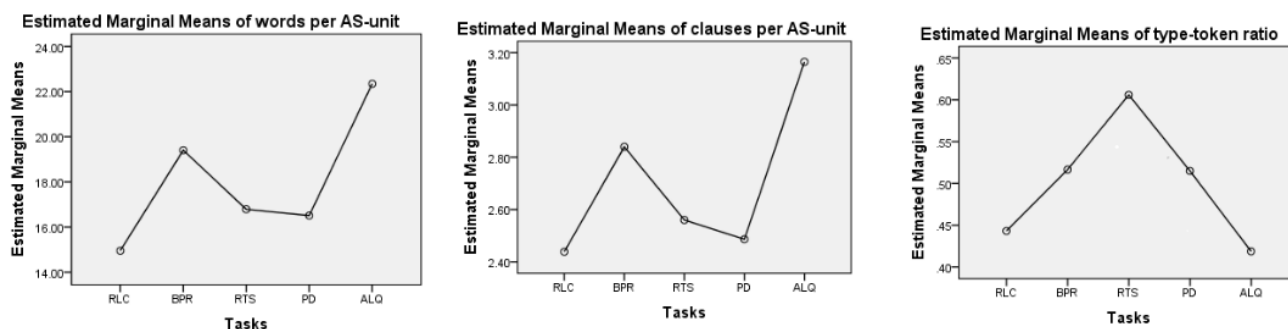
Post Hoc Pairwise Comparisons of Complexity Indices across Test Tasks

Mean Difference		Words per AS-unit		Clauses per AS-unit		Type-token ratio	
		<i>p</i>	Mean Difference	<i>p</i>	Mean Difference	<i>p</i>	
RLC	BPR	-4.452*	.002	-.402*	.042	-.073*	.003
	RTS	-1.835	.141	-.122	.538	-.163*	.000
	PD	-1.556	.222	-.048	.758	-.072*	.004
	ALQ	-7.395*	.000	-.727*	.000	.024	.215
BPR	RLC	4.452*	.002	.402*	.042	.073*	.003
	RTS	2.618	.053	.280	.178	-.089*	.001
	PD	2.897*	.045	.354	.068	.002	.956
	ALQ	-2.943*	.046	-.325	.124	.098*	.000
RTS	RLC	1.835	.141	.122	.538	.163*	.000
	BPR	-2.618	.053	-.280	.178	.089*	.001
	PD	.279	.760	.074	.645	.091*	.000
	ALQ	-5.560*	.000	-.605*	.002	.187*	.000
PD	RLC	1.556	.222	.048	.758	.072*	.004
	BPR	-2.897*	.045	-.354	.068	-.002	.956
	RTS	-.279	.760	-.074	.645	-.091*	.000
	ALQ	-5.839*	.000	-.679*	.000	.096*	.000
ALQ	RLC	7.395*	.000	.727*	.000	-.024	.215
	BPR	2.943*	.046	.325	.124	-.098*	.000
	RTS	5.560*	.000	.605*	.002	-.187*	.000
	PD	5.839*	.000	.679*	.000	-.096*	.000

* Mean difference is significant at .05 level

Figure 4

Comparison of Means of Complexity Indices across Test Tasks



Under the accuracy category, significant effects of testing tasks were found on two indices, while the other one was close to significant. For pronunciation errors per AS-unit, Wilk's Lambda=0.28, $F(4, 33)=21.77$, $p=.000$; for lexical errors per As-unit, Wilk's Lambda=0.55, $F(4, 34)=6.98$, $p=.000$; for syntactic errors per AS-unit, Wilk's Lambda=0.78, $F(4, 34)=2.43$, $p=.066$. Results of post-hoc comparisons are displayed in Table 8, and Figure 5 shows plotted difference of means of all three accuracy indices across the 5 tasks. In the 9 pairs of comparison groups against non-OPI tasks, RLC was significantly lower in value (more accurate) in 7 of them, and was insignificantly lower in 2. On the other hand, BPR was significantly lower in 3, insignificantly lower in 1, and insignificantly higher in 5. BPR also elicited significantly less accurate speech samples in all 3 comparison groups against RLC. The fact that RLC induced higher accuracy in all 9 comparison groups than non-OPI tasks, while the case for BPR was very mixed indicates that the OPI elicits more accurate speech samples except for its effective probe phases than non-OPI tasks.

Table 8

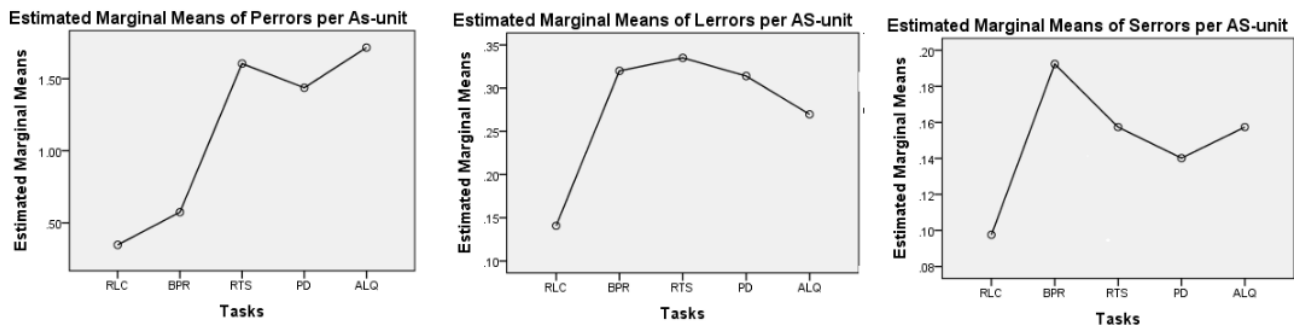
Post Hoc Pairwise Comparisons of Accuracy Indices across Test Tasks

Mean Difference	pronunciation errors per AS-unit		lexical errors per AS-unit		syntactic errors per AS-unit		
	<i>p</i>	Mean Difference	<i>p</i>	Mean Difference	<i>p</i>		
RLC	BPR	-.227*	.005	-.179*	.001	-.095*	.017
	RTS	-1.258*	.000	-.194*	.001	-.060	.123
	PD	-1.090*	.000	-.173*	.002	-.043	.289
	ALQ	-1.370*	.000	-.129*	.022	-.060*	.026
BPR	RLC	.227*	.005	.179*	.001	.095*	.017
	RTS	-1.031*	.000	-.015	.829	.035	.497
	PD	-.863*	.000	.006	.920	.052	.128
	ALQ	-1.142*	.000	.050	.412	.035	.360
RTS	RLC	1.258*	.000	.194*	.001	.060	.123
	BPR	1.031*	.000	.015	.829	-.035	.497
	PD	.168	.307	.021	.738	.017	.719
	ALQ	-.111	.541	.065	.381	-0.00	.999

PD	RLC	1.090*	.000	.173*	.002	.043	.289
	BPR	.863*	.000	-.006	.920	-.052	.128
	RTS	-.168	.307	-.021	.738	-.017	.719
	ALQ	-.279*	.024	.044	.452	-.017	.627
ALQ	RLC	1.370*	.000	.129*	.022	.060*	.026
	BPR	1.142*	.000	-.050	.412	-.035	.360
	RTS	.111	.541	-.065	.381	0.00	.999
	PD	.279*	.024	-.044	.452	.017	.627

* Mean difference is significant at .05 level

Figure 5
Comparison of Means of Accuracy Indices across Test Tasks



Under the fluency category, significant effects of testing tasks were found on two indices, while the other one was insignificant. For total syllables per minute, Wilk’s Lambda=0.66, $F(4, 34)=4.31, p=.006$; for mean length of run, Wilk’s Lambda=0.81, $F(4, 36)=2.05, p=.108$; for pauses per minute, Wilk’s Lambda=0.70, $F(4, 34)=3.66, p=.014$. Results of post-hoc comparisons are displayed in Table 9, and Figure 6 shows plotted difference of means of all three fluency indices across the 5 tasks. In the 9 pairs of comparison groups against non-OPI tasks, RLC induced significantly more fluent speech samples in 4 of them, insignificantly more fluent ones in the other 5. On the other hand, BPR was also significantly more fluent in 5 groups, and insignificantly more fluent in 3. There was no significant difference when RLC was compared to BPR. We can then conclude that OPI in general elicited more fluent speech samples than non-OPI tasks, including its effective probe phases.

Table 9
Post-Hoc Pairwise Comparisons of Fluency Indices across Test Tasks

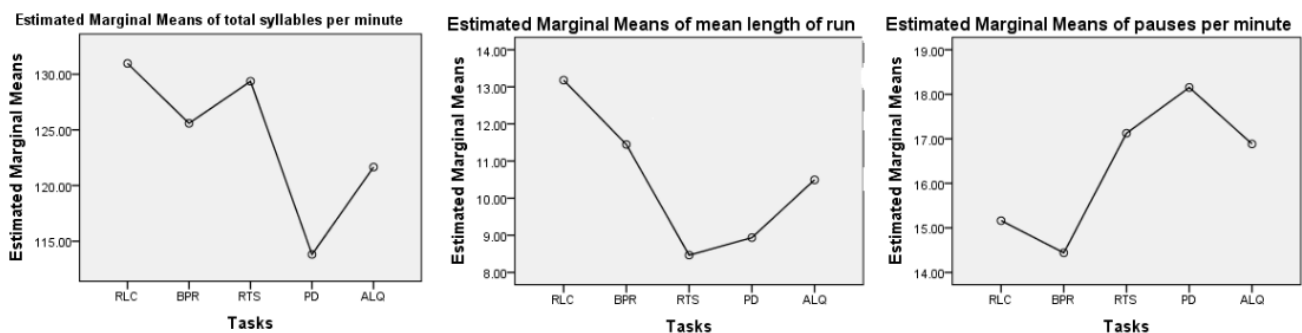
Mean Difference		Total Syllables per Minute		Mean Length of Run		Pauses per Minute	
		<i>p</i>	Mean Difference	<i>p</i>	Mean Difference	<i>p</i>	
RLC	BPR	5.381	.271	1.735	.309	.723	.500
	RTS	1.610	.926	4.716*	.023	-1.963	.081
	PD	17.163*	.000	4.244	.071	-2.991*	.008
	ALQ	9.313*	.039	2.686	.086	-1.721	.129

BPR	RLC	-5.381	.271	-1.735	.309	-.723	.500
	RTS	-3.771	.824	2.981*	.014	-2.686*	.025
	PD	11.782*	.009	2.510	.105	-3.714*	.001
	ALQ	3.932	.332	.951	.309	-2.444*	.021
RTS	RLC	-1.610	.926	-4.716*	.023	1.963	.081
	BPR	3.771	.824	-2.981*	.014	2.686*	.025
	PD	15.553	.360	-.471	.718	-1.027	.291
	ALQ	7.703	.637	-2.030	.105	.242	.823
PD	RLC	-17.163*	.000	-4.244	.071	2.991*	.008
	BPR	-11.782*	.009	-2.510	.105	3.714*	.001
	RTS	-15.553	.360	.471	.718	1.027	.291
	ALQ	-7.850*	.032	-1.559	.210	1.270	.059
ALQ	RLC	-9.313*	.039	-2.686	.086	1.721	.129
	BPR	-3.932	.332	-.951	.309	2.444*	.021
	RTS	-7.703	.637	2.030	.105	-.242	.823
	PD	7.850*	.032	1.559	.210	-1.270	.059

* Mean difference is significant at .05 level.

Figure 6

Comparison of Means of Fluency Indices across 5 Tasks



To sum up the above findings, we argue that hypothesis 3 was validated with a condition added, that is: The speech samples elicited in ACTFL OPI are less complex, more accurate and fluent in comparison to those elicited in non-OPI oral test tasks in general, except for those from the effective probe phases.

To address hypothesis 4 on learner perceptions favouring OPI or not, paired sample t-tests were run between responses to the eight pairs of questions (Q1 to Q16) in the survey (Appendix 3). As shown in Table 10, all results favored OPI except for pair 5 (motivation), which did not show significant difference between the two tests. Pair 1 showed participants thought OPI more faithfully reflected their oral proficiency than non-OPI; pair 2 showed participants favored the experience of taking OPI than non-OPI; pair 3 showed participants thought they gave better performance in OPI than non-OPI; pair 4 showed participants thought they gained more confidence of learning after taking OPI than non-OPI; pair 6 showed the participants felt less anxious after taking OPI than non-OPI; pair 7 showed the participants

thought that OPI more authentically reflected real-life language use than non-OPI; pair 8 showed the general acceptance of OPI was much higher than non-OPI.

Table 10
Paired-sample T-test Scores for Survey Questions

		Mean	N	SD	t	p
Pair 1	Q1 (OPI)	4.20	40	.648	4.363	.000
(faithfully reflecting proficiency)	Q2 (Non-OPI)	3.38	40	.979		
Pair 2	Q3 (OPI)	4.33	40	.656	5.312	.000
(liking of the test experience)	Q4 (Non-OPI)	3.38	40	.774		
Pair 3	Q5 (OPI)	4.18	40	.984	7.127	.000
(perceived better test performance)	Q6 (Non-OPI)	2.33	40	.944		
Pair 4	Q7 (OPI)	3.60	40	.900	6.259	.000
(getting more confidence)	Q8 (Non-OPI)	2.55	40	.749		
Pair 5	Q9 (OPI)	3.90	40	.871	1.325	.193
(getting more motivation)	Q10 (Non-OPI)	3.65	40	.975		
Pair 6	Q11 (OPI)	2.28	40	.905	-4.443	.000
(getting more anxiety)	Q12 (Non-OPI)	3.13	40	1.202		
Pair 7	Q13 (OPI)	4.15	40	.864	6.548	.000
(reflecting real life use)	Q14 (Non-OPI)	2.95	40	.815		
Pair 8	Q15 (OPI)	4.18	40	.874	4.707	.000
(better acceptance in general)	Q16 (Non-OPI)	3.03	40	.920		

The above results not only provided perceptual evidence to support hypothesis 1 (pair 1 and 3) and hypothesis 3 (pair 7), but also supported hypothesis 4 from different angles. First, participants considered the OPI experience more authentic, in terms of better reflecting real life language use (pair 7). Student experiences of OPI were also more positively viewed (pair 2 and 8), and better ranked in terms of affective factors (pair 4: stronger motivation; pair 6: less anxiety). Such positive views of OPI compared to non-OPI tests can be seen as justifying the higher cost of OPI. We also argue for positive washback effects by building communicative language practices into second language curricula more broadly, connecting the experience of language testing to learning skills which are more relevant for use in real life, than on curricula based on traditional grammar-based knowledge.

To gain a more insightful understanding of potential positive washback effects, we analysed the open question and interview data for key themes that were mentioned frequently. As summarized in Table 11, the most frequent themes found in the interview and open question data were consistent with those found in the survey. Comments that the OPI “boosts confidence”, “encourages speaking”, “tests language use (instead of linguistic knowledge)” and was preferred by participants to non-OPI tests are in line with the goals that the ACTFL OPI was supposed to achieve. The cross-validation of survey and interview findings thus confirmed hypothesis 4.

However, we found some unexpected feedback from the interview data. Participants considered OPI put them in a “comfort zone”, while the non-OPI test pushed them to their limit and highlighted some weaknesses. The notion of “comfort zone” could work against the design goals of OPI, being somewhat two-sided. It could mean both offering a better experience, and also entailing the inability to differentiate participants, an aspect of our findings which we discuss further below.

Table 11
Key Themes Emerging from Group Interviews

Thematic code	Frequency of references
I like OPI	8
OPI tests language use	7
HSKK tests knowledge	5
OPI puts me in the comfort zone	5
OPI is adaptive	5
HSKK pushes me to limit	4
OPI boosts confidence	4
HSKK more difficult	3
HSKK made me less confident	2
HSKK spots weakness	2
OPI is weak in differentiating	2
OPI gives more freedom	2
OPI is more fun	2
HSKK is more comprehensive	1
HSKK is more restrictive	1
HSKK is rigid	1
HSKK is less humanized	1
I like HSKK	1
OPI encourages speaking	1

5 Discussion

In this study we used a number of carefully designed measures to compare oral performance in OPI to non-OPI tests among second language Chinese learners at different levels of exposure to Chinese in an immersion setting in China, and to gather data on testee perceptions of the two types of tests, in relation to promoting real-life authentic communication skills. To evaluate the claimed advantages of OPI against its resource-intensive demands, we tested four hypotheses: 1) that OPI ratings would be more dispersed if using OPI rubrics, 2) that OPI speech samples would be more dispersed, due to the proficiency-adaptive test design, 3) that OPI speech would be more less complex but more accurate and fluent as unprepared dialogic speech, and 4) that participants would perceive OPI as more related to authentic communication. In general, the data outcomes as shown above can be taken to indicate evidence in favour of the claimed benefits of OPI. Linguistic performance benefits were clearly found when using objective CAF measures of complexity, accuracy and fluency (as predicted in hypothesis 3). In our data we found that the OPI's benefit for accuracy did not extend to the effective probe phase, but we do not see this as particularly significant. The findings for that phase comprised only one sub-component of all the CAF analyses; furthermore, mixed outcomes for CAF interactions have regularly been found in the literature (e.g. [Awwad et al, 2017](#)); however, it is an aspect that would merit further exploration, particularly cross-linguistically, to validate our findings here for Mandarin against data from other languages.

There was a less clear outcome in regard to our first hypothesis about the value of OPI to create clearly dispersed ratings; as noted above, this was only partially supported, which we now discuss further.

While OPI did not significantly disperse scores overall, dispersion was better for OPI than all the three core tasks in non-OPI when considering performance during the probe phases (hypothesis 1), and when taking into account the objective linguistic indicators of lexicogrammatical knowledge and complexity, accuracy and fluency in production (hypothesis 2). However, OPI dispersion was weaker in the same dimension if subjectively rated, whether using HSKK or IELTS style ratings. We suggest this relates to fundamental differences in the two tests' grading criteria. In its official familiarization manual, the ACTFL OPI overtly claims to be an integrative assessment that evaluates spoken language ability from a global perspective rather than based on the presence or absence of any given linguistic feature. Linguistic components are viewed in the context of their contribution to overall speaking performance. (ACTFL, 2012). A summary of assessment criteria attached to the manual (Appendix 5) specified the above principles. By contrast, the IELTS speaking band descriptor uses highly linguistic indicators (Appendix 6), while the HSKK rating primarily focuses on the testee's ability to complete the test task or not.

We suggest these contrastive rubrics led to the difference in grading. To verify this, we reinspected the correlation coefficients across test scores and between test scores and learners' linguistic component scores. As shown in Table 12, all three test scores were highly correlated. The HSKK score and IELTS score were almost collinear ($r=.989$, $p<.001$), which was not surprising, since they evaluated the same set of oral data. However, other correlations were rather different. The correlation coefficient between OPI score and vocabulary score ($r=.494$, $p<.001$) and between OPI score and grammar score ($r=.385$, $p<.005$) were lower than those found for HSKK score with vocabulary ($r=.585$, $p<.001$) or with grammar ($r=.424$, $p<.001$); the lower correlation coefficients was also found for IELTS score and vocabulary ($r=.622$, $p<.001$) and grammar ($r=.419$, $p<.001$). These findings were within our expectation, and confirm our conclusion that shifting from measuring specified linguistic components to emphasizing "overall" performance was the cause of OPI's inability to show dispersion across the testees. However, it should also be noted that Language Testing International (LTI, the testing body licensed by ACTFL) now provides diagnostic grids upon request, which deliver linguistically-centered feedback on what is lacking to reach the next level. We believe this instrument, if integrated in some way into ratings, could provide added data for dispersing testee ratings.

Table 12

Correlations Between Oral Test Ratings and Linguistic Component Scores

	OPI score	HSKK score	IELTS score	vocabulary test score	grammar test score
OPI score	1	.744**	.745**	.494**	.385*
HSKK score		1	.989**	.585**	.424**
IELTS score			1	.622**	.419**
vocabulary test score				1	.539**
grammar test score				*	1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Based on the above findings, the design of the ACTFL OPI and its effectiveness can be re-evaluated. When compared with typical non-OPI tasks, the current OPI test did realize three of its expected advantages, which were eliciting more linguistically-dispersed speech samples, eliciting more real life-like speech samples (though the two happened in different phases of the test) and better acceptance

by testees, which might potentially promote their learning. Moreover, the effect of eliciting more linguistically-dispersed language samples was achieved while participants felt they were within their comfort zone. This reflects the OPI's aim to subtly conceal "the phases that push the participants to their limit (probes) by making them feel at ease with its gradual approaching elicitations (including the warm up and wind down)" (ACTFL, 2012). This also means the high cost and complex procedures of the OPI were rewarded with both differentiating effectiveness and good participant experience. However, the OPI might be comparatively weaker in differentiating reported test outcomes, and might offset its advantage in eliciting linguistically dispersed language samples, which can be arduous to achieve through the test's costly and complex design.

Overall, we believe the OPI reliably retains the requirement of being a useful test, as found in previous research (Henning, 1992; Surface & Dierdorff, 2003). Referring back to Bachman & Palmer's (1996) original formula for checking usefulness, Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality (p.18). In the current study, we found that OPI excels in Authenticity, Interactiveness and Impact, in comparison to non-OPI tests. The test may be challenging in terms of Practicality, by requiring more resources to train raters and conduct the tests, but the benefits can outweigh these challenges. We also suggest that Construct validity could be improved by adding information to reflect participants' abilities on specified linguistic components, but this can be achieved by attaching the diagnostic grids issued by LTI to OPI raw scores.

6 Conclusion

This study is, to our knowledge, the first to make a comprehensive linguistically-informed comparison of oral outputs for Chinese second language learners in two different test formats, namely ACTFL OPI and HSKK. Our study thus aimed to inform current theoretical and institutional debates about the impact and use of high-stakes tests of oral proficiency as well as what constitutes evidence of oral proficiency for learners of Mandarin, one of the fastest growing second languages globally. We conclude that the current form of the ACTFL OPI test remains a reliable, practical, positively regarded test of authentic proficiency despite its cost disadvantages. Compared to HSKK, as a non-OPI format test, the OPI elicited generally more dispersed as well as more real life-like and communicative speech samples, particularly when analysed through objective linguistic components of complexity, fluency and accuracy in use of vocabulary and grammar. The test was also perceived by participants as representing more lifelike "language usage" compared with non-OPI tests, creating a potential positive washback effect. However, the apparent linguistic dispersion effect in generating a good range of language performances in our study seemed not to lead to well-dispersed final test outcomes. We suggest this problem could be solved by adding the diagnostic grids issued by Language Testing International, the testing body licensed by ACTFL. We consider the importance of this step should not be under-estimated, given the use of OPI as such a high-stake oral proficiency test. Without it, the high cost of OPI's tester training, test delivering and grading procedures may not always appear to be worth investing in. We call for further research on the practical and institutional implications of adding such data to ACTFL testing and results processes.

We acknowledge the limitations of the current study which only included data gathered from Chinese second language learners. We do not believe that the type of target languages would affect the results of the study, particularly given the claimed universality of the ACTFL OPI for covering all languages. We call for replications of this study, which should be reasonably easy to do, in view of the range of non-OPI tests available as counterparts to OPI for most commonly taught second languages, as outlined in our introduction. If the findings of this study were cross-validated, it would be safe to claim that the ACTFL OPI can indeed achieve the full ambition of its design goals with only a little more emphasis on including data from the more linguistically focused diagnostic grids.

Appendix

Appendix 1

The Phases of the OPI

The Warm Up	Iterative Process		The Wind Down
	The Level Checks	→The Probes	
This first phase of the OPI serves as the introduction to the interview. It consists of greetings, informal exchanges of pleasantries, and conversation openers pitched at a level which appears comfortable for the speaker. Every OPI begins with the assumption that a conversation will take place (Intermediate Level).	When the speaker has settled into the interview and appears to be reasonably comfortable using the target language, the interviewer moves to the next phase of the OPI, the level checks. The interviewer engages the speaker in conversation on several topics of interest so that the tasks characterizing any given level can be performed. Level checks are questions that elicit the performance floor, the linguistic tasks, and contexts of a particular level which can be handled successfully.	Once the interviewer has begun to establish that the speaker can handle the tasks and topics of a particular level, the interview proceeds to the next phase, the probes. The purpose of the probes is to discover the ceiling or limits of the speaker's proficiency, i.e., the patterns of weakness. This is done by raising the level of the interview to the next higher major level in an attempt to discover the level at which the speaker can no longer sustain functional performance.	The final phase returns the speaker to a comfortable level of language exchange and ends the OPI on a positive note.
↖ The Role Play ↗			
A transactional or social situation can serve as either an additional level check or probe as needed in a particular interview.			

(ACTFL, 2012)

Appendix 2

The structure of the Non-OPI test

Tasks	Description	Number/difficulty of items	Preparation time (minutes)	Allowed run time (minutes)
T1	repeating sentences after listening	10 (5 primary, 5 intermediate)	0	3
T2	answering short questions	10 (primary)	0	3

	T3	retelling stories	2 (advanced)	0	6
core tasks	T4	Describing a picture after preparation	2 (intermediate)	5	4
	T5	answering questions written on the paper after preparation	3 (1 for each level)	5	6
	T6	oral reading	1 (advanced)	5	3
	total number	6	28	15	25

Appendix 3

Survey: Perceptions of Oral Tests

You have taken both Oral Proficiency Interview (OPI) and conventional oral performance test (HSKK), and now we have some questions concerning these two types of tests. Please circle the number that best reflects your feeling, with 1 being strongly disagree, 5 being strongly agree.

1. I think OPI can faithfully reflect my oral proficiency level.

1 2 3 4 5

2. I think HSKK can faithfully reflect my oral proficiency level.

1 2 3 4 5

3. I like the experience of taking OPI test.

1 2 3 4 5

4. I like the experience of taking HSKK test.

1 2 3 4 5

5. I think I gave better performance in OPI test.

1 2 3 4 5

6. I think I gave better performance in HSKK test.

1 2 3 4 5

7. I think I have stronger confidence in speaking Chinese after taking OPI test.

1 2 3 4 5

8. I think I have stronger confidence in speaking Chinese after taking HSKK test.

1 2 3 4 5

9. I felt more motivated to learn Chinese after taking OPI test.

1 2 3 4 5

10. I felt more motivated to learn Chinese after taking HSKK test.

1 2 3 4 5

11. I felt more anxious in learning Chinese after taking OPI test.

1 2 3 4 5

12. I felt more anxious in learning Chinese after taking HSKK test.

1 2 3 4 5

13. I think the content of OPI test can reflect real life language use.

1 2 3 4 5

14. I think the content of HSKK test can reflect real life language use.

1 2 3 4 5

15. Overall, I like the form of OPI test.

1 2 3 4 5

16. Overall, I like the form of HSKK test.

1 2 3 4 5

17. Please write your comment (advantage/insufficiency) on OPI test:

18. Please write your comment (advantage/insufficiency) on HSKK test:

Appendix 4

Group interview questions

1. Between the OPI and HSK, which test do you like better? Why?
2. Which test provide better experience for participants? Why?
3. Which parts do you like in the OPI test, which parts do you dislike? Why?
4. Which parts do you like in the HSKK test, which parts do you dislike? Why?
5. Which test do you think can reflect your actual proficiency? Why?
6. Which parts of OPI should be changed if it wants to improve?
7. Which parts of HSKK should be changed if it wants to improve?
8. Are there any other thoughts on the two tests to share?

Appendix 5

A Summary of Assessment Criteria for the ACTFL OPI

Proficiency level	Global Tasks and Functions	Context / Content	Accuracy	Text Type
Superior	Discuss topics extensively, support opinions and hypothesize. Deal with a linguistically unfamiliar situation.	Most formal and informal settings. Wide range of general interest topics and some special fields of interest and expertise.	No pattern of errors in basic structures. Errors virtually never interfere with communication or distract the native speaker from the message.	Extended discourse
Advanced	Narrate and describe in major time frames and deal effectively with an unanticipated complication.	Some informal settings and a limited number of transactional situations. Predictable, familiar topics related to daily activities.	Understood, with some repetition, by speakers accustomed to dealing with non-native speakers.	Paragraphs

Intermediate	Create with language, initiate, maintain, and bring to a close simple conversation by asking and responding to simple questions.	Some informal settings and a limited number of transactional situations. Predictable, familiar topics related to daily activities.	Understood, with some repetition, by speakers accustomed to dealing with non-native speakers.	Discrete sentences
Novice	Communicate minimally with formulaic and rote utterances, lists, and phrases	Most common informal settings. Most common aspects of daily life.	May be difficult to understand, even for speakers accustomed to dealing with non-native speakers.	Individual words and phrases

(ACTFL, 2012)

Appendix 6

IELTS Speaking: Band Descriptors

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar speaks coherently with fully appropriate cohesive features develops topics fully and appropriately	uses vocabulary with full flexibility and precision in all topics uses idiomatic language naturally and accurately	uses a full range of structures naturally and appropriately produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	uses a full range of pronunciation features with precision and subtlety sustains flexible use of features throughout is effortless to understand
8	speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language develops topics coherently and appropriately	uses a wide vocabulary resource readily and flexibly to convey precise meaning uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies uses paraphrase effectively as required	uses a wide range of structures flexibly produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	uses a wide range of pronunciation features sustains flexible use of features, with only occasional lapses is easy to understand throughout; L1 accent has minimal effect on intelligibility

7	<p>speaks at length without noticeable effort or loss of coherence</p> <p>may demonstrate language-related hesitation at times, or some repetition and/or self-correction</p> <p>uses a range of connectives and discourse markers with some flexibility</p>	<p>uses vocabulary resources flexibly to discuss a variety of topics</p> <p>uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices</p> <p>uses paraphrase effectively</p>	<p>uses a range of complex structures with some flexibility</p> <p>frequently produces error-free sentences, though some grammatical mistakes persist</p>	<p>shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8</p>
6	<p>is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation</p> <p>uses a range of connectives and discourse markers but not always appropriately</p>	<p>has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inaccuracies, generally paraphrases successfully</p>	<p>uses a mix of simple and complex structures, but with limited flexibility</p> <p>may make frequent mistakes with complex structures though these rarely cause comprehension problems</p>	<p>uses a range of pronunciation features with mixed control</p> <p>shows some effective use of features but this is not sustained</p> <p>can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times</p>
5	<p>usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going</p> <p>may over-use certain connectives and discourse markers</p> <p>produces simple speech fluently, but more complex communication causes fluency problems</p>	<p>manages to talk about familiar and unfamiliar topics but</p> <p>uses vocabulary with limited flexibility</p> <p>attempts to use paraphrase but with mixed success</p>	<p>produces basic sentence forms with reasonable accuracy</p> <p>uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems</p>	<p>shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6</p>

4	cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice rarely attempts paraphrase	produces basic sentence forms and some correct simple sentences but subordinate structures are rare errors are frequent and may lead to misunderstanding	uses a limited range of pronunciation features attempts to control features but lapses are frequent mispronunciations are frequent and cause some difficulty for the listener
3	speaks with long pauses has limited ability to link simple sentences gives only simple responses and is frequently unable to convey basic message	uses simple vocabulary to convey personal information has insufficient vocabulary for less familiar topics	attempts basic sentence forms but with limited success, or relies on apparently memorised utterances makes numerous errors except in memorised expressions	shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	pauses lengthily before most words little communication possible	only produces isolated words or memorised utterances	cannot produce basic sentence forms	Speech is often unintelligible
1	no communication possible no rateable language			
0	does not attend			

(Retrieved from www.ielts.org)

References

- American Council on the Teaching of Foreign Languages (2012). *ACTFL proficiency guidelines*. American Council on the Teaching of Foreign Languages.
- American Council on the Teaching of Foreign Languages (2012). *Oral proficiency interview: Familiarization manual*. Retrieved from: <https://community.actfl.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=1543bfc6-e597-432d-ae6-19bdb7fffeef>
- American Council on the Teaching of Foreign Languages (2017). *OPI Tester Certification Information Packet*. Retrieved from: <https://www.actfl.org/sites/default/files/assessments/ACTFL%20OPI%20Tester%20Certification%20Packet%202018.pdf>.
- American Council on the Teaching of Foreign Languages (2018). *ACTFL OPI examinee handbook*. Retrieved from: <https://www.languagetesting.com/pub/media/wysiwyg/manuals/opi-examinee-handbook.pdf>
- Awwad, A., Tavakoli, P., & Wright, C. (2017). "I think that's what he's doing": Effects of intentional reasoning on second language (L2) speech performance. *System*, 67, 158-169.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), 201-219.
- Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, 36(4), 498-506.
- Everitt, B., & Skrondal, A. (2002). *The Cambridge dictionary of statistics* (Vol. 106). Cambridge University Press.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354-375.
- Fulcher, G. (2014). *Testing second language speaking*. Routledge.
- Halleck, G. B. (1992). The oral proficiency interview: Discrete point test or a measure of communicative language ability? *Foreign Language Annals*, 25(3), 227-31.
- Henning, G. (1992). The ACTFL oral proficiency interview: Validity evidence. *System*, 20(3), 365-372.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461-473.
- IELTS website (2016). IELTS Speaking band descriptors. Retrieved from www.ielts.org
- Isbell, D. and Winke, P. (2019). ACTFL Oral Proficiency Interview – computer (OPIc). *Language Testing*, 36 (3), 467-477.
- Johnson, M. (2000). Interaction in the oral proficiency interview: Problems of validity. *Pragmatics*, 10(2), 215-231.
- Liskin-Gasparro J. E. (2003). The ACTFL Proficiency Guidelines and Oral Proficiency Interview: A brief history and analysis of their survival. *Foreign Language Annals*, 36(4), 484-490.
- Malone, M., & Montee, M. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass* 4(10), 972-986.
- Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45(3), 241-259.
- Omaggio, A. (1986). *Teaching language in context*. Heinle & Heinle.
- Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing*, 17(3), 289-310.
- Shi, J. W. (1998). WaiGuo liu xue sheng 22 lei xian dai han yu ju shi de xi de shun xu yan jiu (A study on the acquisition order of 22 types of syntactic patterns of Mandarin Chinese by CFL learners). *Chinese Teaching in the World*, 4, 77-98.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93-120.
- Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, 20(3), 347-364.
- Surface, E. A., & Dierdorff, E. C. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals*, 36(4), 507-519.
- Swender, E., & Vicars, R. (2012). *ACTFL oral proficiency interview tester training manual*. American Council on the Teaching of Foreign Languages.
- Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, 54(2), 133-150.

- Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals*, 49(1), 75-92.
- Wang, J., An, N., & Wright, C. (2018). Enhancing beginner learners' oral proficiency in a flipped Chinese Foreign Language classroom. *Computer-Aided Language Learning* 31(5-6), 490-521. <https://doi.org/10.1080/09588221.2017.1417872>
- Wright, C. (2020). Effects of task type on L2 Mandarin fluency development. *Journal of Second Language Studies*, 3(2), 157-159. <https://doi.org/10.1075/jsls.00010.wri>.
- Wright, C., Lu, Y., Zhang, J., Zhang, L., & Zheng, Y. (2022). Tests of learning or testing for learning? An exploratory study of motivation and language learning strategies among HSK level 1-3 test-takers in UK. *International Journal of Chinese Language Teaching*, 3(3), 1-19. <https://doi.org/10.46451/ijclt.2022.03.01>.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24(1), 1-27, <https://doi.org/10.1093/applin/24.1.1>
- Zhang, D., & Lin, C-H. (2017). *Chinese as a second language assessment*. Springer.
- Zhang, F., & Du, Y. (2023). Students' willingness to communicate in the online synchronous one-on-one foreign language classroom. *International Journal of Chinese Language Teaching*, 4(3), 94-108. <https://doi.org/10.46451/ijclt.20230306>

Jun Wang is Professor at School of Humanities at Shanghai Jiao Tong University. His research interests include second language acquisition and teaching Chinese as a second language.

Clare Wright is Professor of Linguistics and Language Teaching at the University of Leeds. She works on how linguistic, cognitive and intercultural factors impact on the student experience in globalised higher education; her current research includes projects on acquisition of second-language Mandarin, and models of interactional strategies to support fluency development in different task contexts. Publications include articles with *Journal of Second Language Studies*, *Computer-Aided Language Learning*, as well as a monograph with Cambridge University Press.

Na An is Associate Professor at School of Humanities at Shanghai Jiao Tong University. Her research interests include corpus linguistics and teaching Chinese as a second language, with particular interests in vocabulary acquisition in Business Chinese as Second Language.

Sicheng Wang is a Lecturer at the University International College (UIC), Macau University of Science and Technology. Her research focuses on Teaching Chinese as a Foreign Language, with particular interests in reading and vocabulary acquisition in Chinese as a Second Language.

OPI 有用吗？对比 L2 汉语学习者在 ACTFL OPI 和 HSKK 口语能力测试中的表现和感知

王骏

上海交通大学，中国

克莱尔·赖特

利兹大学，英国

安娜

上海交通大学，中国

王思程

澳门科技大学，中国

摘要

ACTFL 口语能力面试 (OPI) 测试被广泛用于第二语言评估, 声称其优点在于能够引导出流利、真实的对话性语言, 并促进激励性的学习体验。然而, 研究人员尚未就其实用性以及是否具备可操作性达成共识 (Bachman & Palmer, 1996), 特别是在能够直接比较考生在 ACTFL OPI 和非面试类型 L2 口语测试 (例如英语的 IELTS 或汉语的 HSKK) 中的表现和感知方面, 留下了需要填补的空白。在非 OPI 评估中, 英语测试如 IELTS 已被广泛评估, 但 HSKK 仍未得到充分研究, 尽管 L2 汉语学习者数量迅速增长, 这表明需要更深入地调查不同形式如何影响 L2 汉语考生的口语能力及其对测试的态度。本研究旨在通过一种新颖的混合方法案例研究, 填补这些空白, 调查 ACTFL OPI 和 HSKK 在考生表现、语言知识的熟练和真实使用以及测试体验感知方面的可比性。参与者为四十名来自不同母语背景和熟练程度的汉语二语学习者, 之前对这两种测试均不熟悉。每位参与者在为期两天的测试期间, 由专业测试人员使用标准的 ACTFL OPI 和改编的 HSKK 程序进行测试。分析语音数据以评估 a) 评分的离散度, b) 表现的熟练度, 使用复杂性、准确性和流利性的客观指标, 以及“真实生活”中真实语言的证据。此外, 还进行了问卷调查和小组访谈, 以收集参与者对测试体验的感受和看法。结果证实, ACTFL OPI 评分在考生中的离散度不如非 OPI 评分高; ACTFL OPI 语音样本引导出了更多样的语言熟练度和更真实的语言风格; ACTFL OPI 也被认为更积极, 具有潜在的反拨效应, 促进更有动力的学习体验。本研究揭示了使用 ACTFL OPI 相对于非 OPI 测试的局限性和优势, 特别是对于 L2 汉语; 总体结果呼应了当前教育学动向, 鼓励更多真实的交际语言能力, 无论是用于评估还是在实际生活中。

关键词

ACTFL OPI, 口语测试有效性, 汉语, HSKK

王骏，上海交通大学人文学院的教授。研究兴趣包括二语习得和国际中文教育。

克莱尔·赖特，利兹大学语言学与语言教学的教授，研究项目包括二语汉语习得和不同情境下互动流利度发展的模型。

安娜，上海交通大学人文学院的副教授，研究兴趣包括语料库语言学和国际中文教育，特别关注商务汉语作为第二语言的词汇习得。

王思程，澳门科技大学国际学院的讲师，研究兴趣为国际中文教育，尤其关注汉语作为第二语言的阅读与词汇习得。