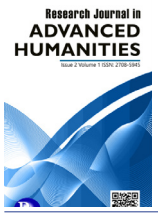

https://doi.org/10.58256/rjah.v4i1.1033

Research Article

Section: Literature, Linguistics & Criticism

Check for updates

Published in Nairobi, Kenya by Royallite Global.

Volume 4, Issue 1, 2023

Article Information

Submitted:23rd November 2022
Accepted: 31st January 2023
Published: 4th February 2023

Additional information is available at the end of the article

ISSN: 2708-5945 (Print)
ISSN: 2708-5953 (Online)

To read the paper online, please scan this QR code

How to Cite:

Jin, T. J., & Ho, I. (2023). Human-machine dialogue modality of English Language oral skills testing among Chinese EFL students. *Research Journal in Advanced Humanities, 4*(1). https://doi.org/10.58256/rjah.v4i1.1033

Human-machine dialogue modality of English Language oral skills testing among Chinese EFL students

Tan Jin Jin[1] & Imran Ho[1*]
[1]Universiti Kebangsaan Malaysia, Malaysia.
Corresponding author: imranho@ukm.edu.my
https://orcid.org/0000-0002-6761-7633

Abstract

With the development of computer technology and the development and application of interactive software, more and more second language oral examinations begin to adopt the form of "human-machine dialogue" as the main way of oral examination. However, the results of these experiments have not been validated in the Chinese context. Therefore, this study aims to identify the affecting factors of students' performance in different English oral tests among Chinese EFL learners. The study used an experimental design. This paper designs and implements a simulated spoken English test of human-machine dialogue modality. 5 Chinese undergraduate students majoring in English language studies and 5 non-English majors are recruited to participate in the experiment. Also, in addition, the experiment also recruited 6 teachers with rich experience in teaching English as a second language as examiners. The results show that there are significant differences in the effective speech frequency in the two tests. Pearson value is.004, and reliability r value is R = 0.04, indicating significant reliability. In terms of hesitation, the duration of all subjects in Q1 and Q2 is significantly different (p<.01). Lexical errors and semantic errors were the most occurred mistakes among students. Finally, the subjects showed high level of anxiety. In terms of self-evaluation of speaking ability, only the group of intermediate English learners show significant differences in the self-evaluation of the two experiments test can truly reflect their oral English ability. This study recommends more research on the human-machine dialogue as the subjects in human-machine dialogue modality cannot get real-time feedback from the communication object, so the subjects are seldom able to notice and self-correct grammatical errors during the expression process. Fourth, the author analyzes and raises research questions about the mistakes made by the subjects in their oral expressions. The human-machine dialogue modality cannot guide the examinees through real-time communication and stimulate the examinees to express the examinee's grasp of relevant second language knowledge which the examiner wants to test.

Keywords: experimental, experimental design, Human-Machine dialogue, language learning, speaking ability.

Introduction

English language education in China has received increased attention over the last few decades, and thus, many believed that this has resulted in great improvement in learners' English proficiency (Wang, 2008). It is statistically approved that China has by far the most English language learners compared to other countries in other parts of the world. Nevertheless, China's rapid urbanization has resulted in a significant shift. Practitioners who are fluent in English in their specialist professions are in high demand in today's world. Particularly, it indicates that it is important to conduct an English language evaluation to identify professional candidates with the best standard of English language skills in order to assist foreign companies and Chinese firms conducting related businesses in the country's market. Many university graduates in China graduated with very prominent results on their College English Test (CET) (Zheng & Cheng, 2018, Razak et al., 2022). In China, where the practice of language education had primarily concentrated on reading and writing skills, the growth of verbal education only emerged in the 1990s. The Cambridge Business English Certificate (BEC) with a verbal element was the first test introduced in China in the year 1993. The verbal element of the Test for English Majors (TEM) system was introduced in 1994. The National Matriculation English Test - Oral Subtest (NMETOS) was first administered in three Chinese provinces in the year 1995 (Li and Wang, 2000). In 1999, the College English Test - Spoken English Test (CET-SET) was implemented. Language assessment scholars in China have performed a variety of longitudinal comparisons to understand different concerns related to the advancement of verbal evaluations. Li and Wang (2000) focused on the creation of the verbal pre-test of the National Matriculation English Test. They discussed various drawbacks related to oral assessment in China such as a large number of candidates for the examination and China's strict limits on personnel and time assets.

With the development of computer technology and the development and application of interactive software, more and more second language oral examinations begin to adopt the form of "human-machine dialogue" as the main way of oral examination (Huang et al, 2021). The economic advantages of the "human-machine dialogue" modality are clear: test organizers can employ a relatively small number of invigilators to organize large numbers of candidates to take the oral test at the same time, and these invigilators do not need to have relevant English background. In addition, unlike the traditional speaking test modality of examiner/examinee interview, the human-machine dialogue test modality enables examiners to score remotely from any place without having to go to the test site in person. This also greatly improves the scoring efficiency and reduces the scoring cost. Another advantage of the human-machine dialogue test modality is that the examiner can only hear the examinee audio data and can't see the examinees, the examiner is more impossible to know the candidate's nationality, gender, etc. These factors may cause score discrimination (although for trained examiners, the possibility of such discrimination problem is almost zero, but it doesn't guarantee absolute zero). Therefore, the human-machine dialogue test modality can ensure that the scoring is not affected by factors other than the English ability of the examinee to the maximum extent, and improve the fairness of the scoring (Ramanarayanan et al., 2017). However, after all, science and technology has not advanced to realize the real sense of "human-machine dialogue" -- examinee and computer natural discourse communication, so researchers still have doubts about the human-machine dialogue modality of second language oral test, and they think that its biggest defect lies in the unauthenticity of dialogue communication. Some researchers have pointed out that in the real world, it is rare for speakers to give a complete speech on a

given topic within a specified time, and candidates are likely to be unable to adapt to such unnatural oral communication, which may affect their oral performance (Qian, 2010). In addition, since the human-machine dialogue measures unnatural oral performance, skeptics also worry that the results may not be completely equivalent to natural oral second-language ability (Ramanarayanan, 2020). Despite the above inevitable defects in human-machine dialogue modality, in today's expanding market of SLA, the resources of second language spoken test are far behind the growth of test demand. This situation makes it difficult for second language examination institutions to completely abandon the human-machine dialogue modality, and choose the relatively inefficient and costly examiner/examinee dialogue modality and group discussion modality. Therefore, referring to the discussion of candidates' oral performance in human-machine dialogue modality, how their oral performance is affected by the characteristics of human-machine dialogue modality and the evaluation of examiners are prerequisite for improving this kind of modality, and are also one step to improve the effectiveness of this oral test modality (Qian et al., 2020, Alakrash & Razak, 2019). Therefore, this modality can be developed and popularized to meet the market demand of second language oral test.

As for the significance of speaking skills, they face many various problems that hinder learners from mastering them during testing; continuously these reasons tend to subject choice, non-appropriate use of modern teaching methods, too little stated time in teaching them, non-stated bases of assessment, and colloquial usage during teaching and weakness of linguistic outcome AlSaleem, (2018). So, an important question might be proposed; how can English language learners who speak other languages master oral communication skills? At the same point, many English and foreign studies such as Abdel-Hamid (1986) and Abu-Rabia (2001) assured that students' behaviours in speaking skills testing need more attention in various stages. Nonetheless, English oral test modes have been inadequately addressed in the literature, in the Chinese context. There is only scanty research on some aspects of oral test modality in the local contexts in some Chinese universities. These studies, descriptive in nature, were restricted to descriptions of how to improve oral literacies, the design and reliability, and the washback effect of these exams. Furthermore, none of these studies touched on English oral test modes by comparing the current testing modes to investigate the preferred test mode, the factors that affect students' performance and the effect of the test modes namely "Machine-examinee" on students at the university level in Chinese universities. Such an investigation will help to find an effective test modality for EFL Chinese students, which is an area that has remained quite detached from L2 research considerations.

Literature Review

Human-machine dialogue modality is represented by TOEFL

The TOEFL speaking test is a one-person computer test, and the situational communication in the speaking test is low. At the same time, the strict time limit of the examination environment and the serious environment of the examination room will also affect the normal oral communication level of candidates. In addition, TOEFL oral test modality is not conducive to Chinese candidates because Chinese students lack an English language environment, even if they actively prepare for the test, it is still difficult to complete the limits of the "Chinglish" thinking as the interference of the mother tongue language. In the tense test environment, it is more difficult to think carefully and use the formal language structure and expressions as the scoring process will be conducted using a computer (AI analytics) based on setup criteria. With the rapid development and popularization of computer science, man-machine

dialogue is gradually adopted by more and more English tests. Almost all the research supporting human-machine oral communication modality take "economy, convenience and large scale of large quantities" as their main advantages (Alakrash et al. 2022). But some researchers have questioned this model, which is based on new technology. One of the reasons for questioning this modality is the acceptance of the candidate. In Qian et al., (2011) 's research, Qian found that the number of the subjects who strongly preferred human-computer dialogue was significantly less than that of the number who strongly preferred the conversation between the examiner and the examinee.

Theoretical Framework

The Socio-Cultural Theory

The Socio-Cultural Theory (SCT) of second language acquisition is based on the theoretical concepts of Vygotsky, Weitzki and Leondev. These twentieth-century theorists challenged linguists who had overlooked the importance of interaction in language acquisition. The Socio-Cultural Theory of second language acquisition emphasizes the centrality of language as an intermediary artefact. According to the SCT, language acquisition is conversational and learning takes place in positive interactions rather than as a product of sociocultural interactions. The scaffolding is considered key to the dialogue process, in which the teacher helps the acquirer perform functions that he or she could not do on her own. Collaborative dialogues are used to help build knowledge and solve problems in the same acquisition process. Although the teacher's interaction is curriculum goal-oriented, the learning process is conversational. Research has shown that private conversations/presentations support second language acquisition.

Communicative Language Ability

In the 1990s, Bachman, an American applied linguist, put forward the Communicative Language Ability modality(CLA), that is, language competence should include "knowledge of grammatical rules and how to use language to achieve specific communicative purposes; Language use is a dynamic process in which the components of language competence interact "(Bachman, 1990:84). The theoretical model of Communicative Language Ability established by Bachman is by far the most comprehensive and complete theory on language competence (Xu Qiang, 2000). Bachman (1990:84) believed that "linguistic communicative competence is the ability to combine linguistic knowledge with the scene features of language functions to create and interpret meanings. He divides communicative competence into three components: language competence, strategic competence and psychophysiological mechanism. According to Bachman, linguistic competence includes two parts: organizational competence and pragmatic competence (Bachman,1990:84-98), which can be subdivided into smaller categories. Organizational Competence determines how a text is organized, which involves the ability to control the formal structure of the language, generate and identify grammatically correct sentences, understand the subject matter and arrange the sentences in the order of the idioms.

Methodology

The materials for this experiment include a set of spoken English exam questions, a good laptop computer, and good pair of headphones, high performance recording pen, questionnaire completed by the subjects, and a questionnaire completed by the examiners.

The speaking test questions used are taken from parts 1 and 2 of the TOEFL Speaking Test bank. The principle of selecting test questions is to be close to the daily life and personal experience of

the subjects, and to have a certain ideological and dialectical nature, so as to ensure that the subjects will not be unable to give a speech due to unfamiliar topics, or have nothing to say due to the topic is too simple and straightforward. In addition, considering that all the subjects are university students, the topic with certain thinking and complexity and close to the immediate interests of the subjects can best stimulate the desire of expression of the subjects, so as to avoid possible participant fatigue and ensure the accuracy of experimental results.

Although the subjects do not have to listen to any voice prompt during the simulation test, the experiment is equipped with a headset with good sound insulation performance to ensure that the subjects are not disturbed by external factors during the simulation test, and the scene of the human-machine conversation mode oral test is restored as far as possible.

The content of the subject questionnaire is basically similar to that of experiment 1, but it includes the on-site experience of the human-machine conversation oral test, the self-evaluation of the test performance, and the reasons for the preference of human-machine conversation modality. The author can understand the specific experience and acceptance degree of the examinees. The design of the examiner questionnaire is mainly to understand the evaluation criteria of the on-site performance of the subjects. In order to know which skill performance of the examinee takes the most weight in the examiners' subjective judgment, examiners are not provided with any scoring criteria, and examiners are asked to score on a scale of 10, with a minimum of 1, a maximum of 10, and a minimum of 0.5 points. The examiner questionnaire asks the examiner to state the composition of their criteria for scoring candidates, and rank the different criteria according to the weight.

The Experimental Subjects

Five undergraduates majoring in English and 5 non-English majors of Experiment 1 are recruited. All of the experimental subjects have oral English test experience (Cet-4, CET-6, TOEFL and IELTS), and their scores in these tests are at the medium level (IBT score 18-22; IELTS Speaking test score 5-6; CET 4/ 6 Oral English Test grade B or C). The purpose of setting the score range of the spoken English test is to avoid the ceiling effect on the one hand, and to avoid the detection effect due to the low spoken English ability of the test subjects. Specifically, 5 undergraduates majoring in English are classified as advanced Learners group. Five non-English major undergraduate students are considered as intermediate English learners. The grade span of these intermediate English learners is larger than that of English majors, because non-English majors only offer English courses in freshmen and sophomores.

In addition, the experiment also recruited 6 teachers with rich experience as examiners. Three of the examiners are native Chinese speakers and three are native English speakers.

5.2.3 The Experimental Steps

First, the experimental subjects are informed of the "privacy". Second, subjects adjust the volume of headphones, brightness of computer display screen and font size displayed on the screen to ensure that the experimental environment is adjusted for the comfort of subjects. Third, the researcher makes the experiment process clear to the subjects. Fourth, after completing the simulated test of human-machine dialogue, the subjects are asked to listen back to the voice recording of the simulated test, recall the scene of obvious phonetic errors, semantic errors and hesitations, and explain the relevant situation and their own psychological activities when the errors occurred. Fifth, the experimental subjects complete a questionnaire for this modality of oral English test. Sixth, in order to protect the privacy of the examinees and simulate the real exam to the maximum extent, all examiners score the examinees through the live recording and fill in the scoring results in the examiner questionnaire. For easy understanding.

Specifically, each group of subjects is randomly divided into two groups again. Intermediate English learners are divided into: 2 in group A and 3 in group B; Advanced English learners are divided into: 3 in group A and 2 in group B. The two Group As complete topic 1 first and then topic 2; The two Group Bs complete in reverse order.

Findings and Analysis

Analysis of the Quality of Subjects' Oral Performance

Effective Speech Frequency

The effective amount of information and fluency are two assessment points that examiners generally value in speaking test. In the data analysis stage of this experiment, a complete script is formed by dictation of the live recordings of the subjects, and then the data are extracted from the script for quantitative and qualitative analysis.

First of all, the number of effective words in the spoken expression of the subject is counted. Effective words are those lexicon words with meaning. In order to make data comparable, production per second is calculated by dividing the effective number of words by the speaking time (lining in seconds) and defined as the effective speaking frequency. The reason for calculating the effective speaking frequency is that the effective speaking frequency reflects the speaker's fluency, which is a direct reflection of the speaker's language ability.

Table.1 Descriptive Statistics of Effective Speech Frequency of Subjects

	Mean	N	Std Dev	Std. Error Mean
Q1	1.4100	10	0.510882	0.161555
Q2	1.0600	10	0.236643	0.074833

The average frequency of the two questions is different. In order to verify whether the mean difference is statistically significant, the author conducts an intra-group difference value test (Independent T-test). The results of the analysis show that there are significant differences in the effective speech frequency of the same subject when completing the two tests. Pearson value is.004, and reliability r value is R = 0.04, indicating significant reliability. Table 5.11 shows the test results.

Table 5.11 Intra-group Difference Test -- Effective Speech Frequency

	Intragroup Difference							Sig. (2-tailed)
	Mean	Std Dev	Std. Error Mean	95% Confidence Interval of the Difference		t	df	
				Lower	Upper			
Q1-Q2 (Effective Speech Frequency)	0.350	0.291548	0.092195	0.141439	0.558561	3.796	9	0.004

The fact that subjects performed significantly differently on the two test questions suggests two

possibilities. First, the subjects are affected by participant fatigue, resulting in inconsistent performance; second, due to the different nature and content of the test questions, there are differences in the degree of difficulty, thus leading to the differences in the performance of the subjects. Due to the use of the counterbalance study, if experiment fatigue has such a significant effect on the subjects, half of the subjects would have reported that item 1 spoke more effectively than item 2, while the other half would have reported the opposite. However, the statistics as a whole show no significant difference, and the first possibility can be largely ruled out. Then the second possibility is analyzed and discussed. In order to prove whether there is a difference in the difficulty of the test question itself, which causes the significant difference in the performance of the subjects, the author carries out correlation statistical analysis. If it can be confirmed that there is a significant intra-group correlation between the effective speaking frequencies of the two test questions, it indicates that the answer performance of the two test questions is stable, and the difference reflected is not caused by personal factors, but is influenced by external variables (test difficulty). Data from intra-group significance tests confirmed this prediction, with a significance value of P=0.

Another relationship that needs to be demonstrated is the relationship between the second language spoken ability of the subjects and their oral performance in the human-machine dialogue spoken test modality. The discussion in the previous section has proved that advanced English learners score significantly higher than intermediate English learners in this experiment. This section will discuss the relationship from the perspective of the effective speech frequency of the subject. If the analysis of the difference value of effective speaking frequency is also significant, it can be directly or indirectly explained as follows: 1) The human-machine dialogue oral modality can effectively reflect the differences in English ability of examinees; 2) The examiners who participate in the experiment score reliably; 3) The speaker's effective speech frequency can effectively reflect the second language speaking ability.

First of all, the difference degree of the average effective speaking frequency of the two questions is statistically analyzed. Table 5. 13 is the analysis result, showing the significant difference between groups (P=0, P< .05). The reliability of the statistical result is R =.95, indicating that the result is highly reliable.

Table 3 Difference Analysis in Oral Performance between Advanced English Learners and Intermediate English Learners

	t-test for Means of Equality						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
Mean of speech frequency	8.914	8	0.000	0.6700	0.075166	0.496666	0.843334

The Table 3 confirm the three points mentioned above and conform to the expectation of this experiment. However, the above analysis shows that the correlation between the test questions and the personal life of the examinee also affects the expression desire and oral performance of the examinee. Only the average speaking frequency of the test subjects is used as the analysis basis, which cannot reflect the impact of the differences of the test questions on the examinees with different second language abilities.

Therefore, the difference analysis of the speaking frequency of the test subjects is carried out separately. The analysis results are shown in Table 4.

Table 4 Difference Analysis in Oral Performance between Q1 and Q2 for Advanced English Learners and Intermediate English Learners

Effective speaking frequency	t-test for Means of Equality						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
Q1	11.235	8	0.000	0.940	0.083666	0.747066	1.132934
Q2	5.547	4.57	0.094	0.400	0.072111	0.233712	0.566288

The table shows there is a significant difference between advanced English learners and intermediate English learners in the oral performance showing moderate reliability. However, there is no significant difference in their oral performance on Q2. Based on the intra-group difference analysis results of the two test questions in Table 3 and the inter-group difference analysis results of the two test questions in Table 4, it can be inferred that advanced English learners are more susceptible to the degree of familiarity with the test questions, the degree of correlation between the test and the subjects, and the subjects' desire to express the test questions and other non-oral English ability factors.

Hesitation

The subjects' spoken fluency can be measured by another dimension -- hesitation. In statistics, 0.500 seconds is taken as the dividing line of the hesitate, that is, the silence time over 0.500 seconds is defined as the hesitate. The 0.500 second hesitate standard is based on the fact that the silence below 0.500 second may be a natural breath or semantic pause during speech. At the same time, the hesitate of less than 0.500 seconds is almost invisible to the listener from the perspective of hearing perception, and does not affect the smooth progress of communication. According to the above standard calculation, the hesitate performance of the subject is counted. Then, the average of all hesitate duration of the subjects in a single question is processed to facilitate the comparison of differences within and between groups. Table 5.15 (a) shows the statistical results of the pause time after average processing.

According to the findings, the average duration of the subjects' hesitate in Q1 is shorter than that in Q2. The average hesitate length of advanced English learners is shorter than that of intermediate English learners. To verify the statistical significance of this set of data, the author uses the same method with the effective speech frequency analysis, and again analyzes the differences within and between groups. Table 5 shows descriptive statistics of hesitates in Q 1 and Q2 for all subjects.

Table 5 Descriptive Data of Subjects' Hesitates (unit: second)

Hesitate	Mean	N	Std Dev	Std. Error Mean
Q1	1.29950	10	0.129249	0.040872
Q2	1.87980	10	0.291491	0.092177

The descriptive data in Table 5 shows that the hesitate time of all subjects in Q1 (mean =1.230 seconds) is significantly shorter than that in Q2 (mean =1.880 seconds). Then intra-group differences are tested. Table 6 shows the statistical analysis results.

Table 6 Inter-group difference test of subjects' Hesitates

	Intragroup Difference							
Hesitate	Mean	Std Dev	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Q1-Q2	-0.58030	0.268051	0.084765	-0.772052	-0.388548	-6.846	9	0.000

It can be observed from Table 6 that the hesitate duration of all subjects in Q1 and Q2 is significantly different ($p<.01$), and has very high reliability ($r= 0.79$). This suggests that, from a perspective of hesitate, participants' fluency in Q1 is significantly higher than their performance in Q2. As mentioned above, the topic of Q1 is more familiar to the subjects than that of Q2. Therefore, it is speculated that the subjects have more content to say for Q1, the expression time is shorter, and the subjects spend more energy on organizing the English lexical structure. These factors result in shorter hesitates for participants to ponder the following statements.

The above quantitative analysis discusses the situation that the oral performance of the subjects is affected by their English ability and test content in human-machine dialogue modality. However, the data can only reflect the effective speaking frequency and hesitate duration of the subjects, and cannot explain the specific speech quality. In addition to fluency and effective amount of information, another evaluation that examiners generally value is the accuracy of language, including grammar, vocabulary and sentence structure. Therefore, in addition to quantitative analysis, it is also necessary to analyze the errors that affect the accuracy of the oral expression from a qualitative perspective.

Accuracy

Judging from the feedback of the examiners in this experiment, the accuracy of language expression can be divided into vocabulary, sentence structure (including grammar) and logical content. The author believes that mistakes in sentence structure can be further divided into fixed collocation and sentence pattern. Therefore, this section will focus on the qualitative analysis and discussion of vocabulary errors, collocation errors, sentence structure errors and logical errors. Lexical errors in this experiment refer to semantic errors. From the oral expressions of all the subjects, mistakes in vocabulary can be divided into the following types: First, the wrong part of speech. Example (1) lists similar errors. The underline indicates the wrong words, and the right column indicates the correct statements.

Example (1).

Wrong Expression	Correct Expression
The most memory moment	The most memorable moment
a modesty man	a modest man
The economic is developing at a high speed.	The economy is developing at a high speed.

Second, semantic errors. The words are semantically inconsistent with what the context inferred the subjects wanted to express. Example (2) illustrates several cases of this type of error. Through the analysis of the above two types of lexical errors, it can be seen that due to the limitations of various linguistic and non-linguistic factors, the subjects encounter mapping difficulty when they map the semantic concept in their mind to the corresponding English expression in oral expression. That is to say, it is commonly said by learners that they cannot translate the Chinese concept into appropriate English words. Therefore, the subjects can only choose semantically similar but not completely appropriate words to replace. In Example (2), "speak a speech" and "grow my English" are typical examples of mistakes caused by inappropriate synonyms. The part of speech error in Example (1) reflects that the subjects first consider mapping semantic concepts with English semantic content, and then consider their lexical forms inflection based on the root progression of the core semantic content. In Example (1), the author speculates that the generation of part-of-speech errors is due to limited time pressure. The working memory of the subjects cannot meet the demand for cognitive ability in the cognitive processing chain from semantic to formal, so the errors of cognitive process occur, which makes the semantic content cannot be presented in the correct form. In addition, there is an error in Example (2), that is, the vocabulary is semantically and formally inaccurate.

The second type of error is collocation error. Collocation errors refer to the subjects confusing and misusing the collocation of fixed phrases in English. Collocational errors are sometimes not too difficult to understand semantics, but they are syntactically incorrect. Example (3) shows a number of subjects with this type of errors.

Example (3)

Wrong Expression	**Correct Expression**
are awareness of	are aware of
make us crying	make us cry
hundreds of	hundreds of

The lexical collocation errors listed in Example (3) can be seen as the result of the subjects not fully mastering the fixed collocation. In other words, the subjects do not completely internalize the fixed collocation as a whole phrase into their learned intermediate language. According to the theory of working memory, the target word is separated and processed separately in the cognitive processing of the subject because the subject is not yet able to internalize the target language in the form of a whole phrase. As a result, the working memory of the subjects could not meet the cognitive processing of the target phrase after being subjected to external pressures such as time pressure (limited-time expression), content pressure (unfamiliar content) and scene task pressure (oral test), resulting in production failure. Another kind of errors involved in fixed collocations are grammatical collocations, including English quantifiers and fixed collocations of comparative levels.

Example 4: type of mistake made by the subject.

Wrong Expression	**Correct Expression**
more pretty	much prettier
so many, so much of money	so much money

The error in Example (4) shows that it is difficult for the subjects to use English quantifiers and comparative levels. First of all, in the series of words denoting quantity, the difference between countable and uncountable nouns is easy to cause difficulty for Chinese learners. For example, the subject self-corrects "many" in "so many, so much of time", indicating that the subject is aware of the difference between countable and uncountable nouns and quantifiers when describing the concept of "much money". The third notable mistake concerns the use of the comparative form in English. Because of the simplicity and scarcity of morphology of Chinese words, Chinese English learners are naturally prone to confuse in using "more" and suffix "-er" when they encounter English comparative qualifiers that can both represent comparative meanings.

The third type of errors analyzed in this experiment are grammatical structure errors. As the name implies, any errors involving grammatical structures (such as tense problems, subject-verb agreement, etc.) are classified as type errors.

Example (5)

Wrong Expression	**Correct Expression**
the memorable moment I have	the memorable moment I had
I feel surprised at that moment	I felt surprised at that moment
I forget the words	I forgot the words
I am not sure when shall I put it back	I am not sure when I shall put it back

Except for the last case, all the other examples of example (5) are grammatical errors about tenses. In Q1 of this experiment, the subjects were asked to describe the most memorable moment in their life so far and why this decision was important. Inevitably, the subjects needed to use the past tense as the main narrative tense when they completed Q1. However, from the analysis of the recordings, most of the subjects more or less replaced the past tense with the present tense, and none of the subjects took the initiative to self-correct the tense errors. There are two possibilities in analyzing these two phenomena and deriving their causes. First, the subjects did not know that they had to use the past tense of an English verb to describe something that had happened in the past. Second, the subjects knew this grammatical rule, but paid very little attention to it and seldom paid attention to and modified this type of errors in oral expressions.

Anxiety in the Examination Room

The detection results are presented in Table 9. The former is descriptive statistical data, and the latter is test result.

Table 9 Descriptive Data of Questionnaire Reliability Test

Questions	Mean	Std Dev	N
I had a bad heart beat during the exam just now.	3.60	0.966	10
I just became rigid during the exam.	3.70	0.9486	10
My personal emotions did not affect my performance in the exam just now.	3.30	0.674949	10

According to the data shown in Table 9, the average anxiety level of the test subjects in this experiment has exceeded 65% of the five-point scale, which can be considered as a certain degree of anxiety. These data show that the subjects generally feel nervous during the examination.

Second question answers revealed that 60% strongly agreed with the situation described in the test. The author believes that the more strongly the subjects experienced the rigidity of thinking during the examination, the more strongly they felt anxious in the examination room. The authors suggest that this may be due to the intense concentration of participants in the particular state of taking a test, or even to the excitement of completing the test challenge. This experiment also adopts the qualitative method to investigate and analyze the on-site behavior of the subjects' anxiety in the examination room. The specific research method is as follows: after the subjects complete the two simulation test questions, they play back the recording immediately, and guide the subjects to recall the relevant psychological, physiological and emotional changes in the examination process under the prompts of the recording.

Subjects' Self-perception of Second Language Ability

In order to explore the participants' self-perception of their oral English ability, the participants are asked to make a self-evaluation of their oral English ability in a questionnaire survey. At the same time, in order to explore their self-perception of the examination performance in the human-machine dialogue modality, the questionnaire also requires the subjects to make a self-evaluation. The two self-assessments are conducted on a Likert-5, with 1 being the lowest score and 5 being the highest score.

The data shows that the average score of self-assessment of English ability is 2.4 (N=10), and the average score of self-assessment of oral English performance in human-machine dialogue is 2 (N=10). And the correlation between the two groups was significant (Pearson =.002). This indicates that the subjects of this experiment believe that their performance in the human-machine dialogue modality of oral English test can truly reflect their oral English ability. Table 5.21 shows the results of the above statistical analysis.

Table 10 Inter-group Correlation Test of Self-assessment of English Ability and Oral Performance of the Subjects

		N	Correlation	Sig.
subject	self-assessment of English ability &oral English performance	10	0.839	0.002

Secondly, according to their English ability, the subjects are divided into advanced English learners group and intermediate English learners group, and the differences between the groups are tested. It is found that there is no significant difference between the two groups in the scores of self-assessment of English ability and self-assessment of examination performance. The average self-assessment of English proficiency is 2.80 (N =5) for the advanced group and 2.00(N=5) for the intermediate group. The average self-assessment of performance is 2.20 (Number of students =5) for the advanced group and 1.80 (number of students =5) for the Intermediate group. The Pearson value of inter-group difference between the two groups is p = .14 and p=. 20. This indicates that there is no statistically significant difference between the two groups in these two self-assessments, and both groups underestimate their English ability and test performance. Table 5.22 shows the results of the above statistics.

Table 11 Inter-group Difference Test of Self-assessment of English Ability and Oral Performance of the Subject

	t-test of inter-group mean					
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
						Lower Upper
self-assessment of English ability	1.633	8	0.141	0.80	0.489898	-0.329707
self-assessment of oral English performance	1.414	8	0.195	0.40	0.282843	-0.252236

According to the average values of each group in Table 5.23, the self-assessment data of English ability of subjects in Experiment 1 (both advanced group and intermediate group) is generally higher than that of subjects in Experiment 2. To further investigate whether the difference is statistically significant, an inter-group difference test is conducted. Table 5.24 shows the statistical analysis results of the inter-group difference tests.

Table 13 Inter-group Differences of Self-assessment of English Ability (Experiment 1 and Experiment 2)

self-assessment of English proficiency	t-test of intergroup mean						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence interval of the Difference	
						Lower	Upper
advanced	-1.414	8	0.195	-0.60	0.424264	-1.578355	0.378355
intermediate	3.015	13	0.008	-0.20	0.374166	-1.062828	0.662828

The statistical results show that although the self-evaluation of English ability of the two groups of subjects in experiment (1) was more optimistic than that of the corresponding subjects in experiment (2), statistically speaking, only the group of intermediate English learners show significant differences in the self-evaluation of the two experiments (Pearson value is P =.008).

At the logical level, the significant difference can be attributed to two possibilities: first, the subjects in experiment (1) overestimated their English ability while in experiment (2) they correctly assessed their English ability; second, subjects in experiment (2) underestimated their English ability, while self-evaluation in experiment (1) was objective and correct. If it is the former, it indicates that although the examiner/examinee dialogue modality will negatively affect the self-evaluation of the English ability of the subjects, this negative effect is beneficial to the subjects, because it can make them correctly perceive their English ability (rather than overestimate); But if it is the latter, it shows that the negative effect of the test is not conducive to the subjects. Because it will make the subjects feel depressed for their poor English ability, and then affect their learning enthusiasm.

Another important reference that can be used to explore the reasons for the self-assessment differences of English ability between the two groups is the examiners' rating of the oral performance of the subjects in Experiment 2. If the examiners' ratings of the subjects in experiment (2) are significantly higher than the subjects' self-assessment of their English ability, it could be proved that the subjects in experiment (2) underestimate their English ability. It should be noted that since it has been proved that there is no significant difference between the self-assessment of the English ability and the self-assessment of their performance in the examination room, it is possible to compare the self-assessment of the English ability of the subjects with the examiners' assessment. Technically, since the examiner's rating in this study was on a 10-point scale and the subjects' self-assessment is on a 5-point scale, it is necessary to multiply the examiner's rating by 50% in advance for statistical analysis so that the two are comparable. In the statistical analysis, the oral performance of the subjects is taken as an independent variable, and the evaluation of the performance by both native Chinese examiners and native English examiners and the subjects as dependent variables.

Table 14 shows the statistical analysis results after data processing.

		Intergroup Difference							Sig. (2-tailed)
		Mean	Std Dev	Std. Error Mean	95% Confidence Interval of the Difference		t	df	
					Lower	Upper			
G1	NEE-- self-assessment of the subject	1.0170	0.969502	0.306584	0.323460	1.710540	3.317	9	0.009
G2	NCE-- self-assessment of the subject	0.7510	0.730638	0.231048	0.228333	1.273667	3.250	9	0.010

As can be seen, the average scores of both Native English examiners and Native Chinese examiners are significantly higher than the self-assessment of the English ability of the subjects. The Pearson value of the former is P = .009, while that of the latter is P = .010. The reliability indices are r = 0.81 and r = 0.83 respectively, indicating that the statistical analysis results are highly reliable.

The above statistical results confirm the second possibility mentioned above, that is, the subjects in this experiment significantly underestimate their actual oral English ability. This shows that the human-machine dialogue modality of oral test has a significant negative effect on the self-recognition and affirmation of English ability of examinees, and will cause examinees to wrongly underestimate their oral ability of second language. As a result, students will have a sense of psychological frustration, lose confidence in learning, weaken the motivation of learning, and affect the process of second language acquisition

Discussion

Considering that the subjects of this experiment are at least intermediate English learners and the correct use of the past tense of verbs exists in the subjects' oral expressions, the first possibility mentioned above can be excluded. Thus, there is only a second possible explanation, that is, subjects do not pay enough attention to the grammatical phenomenon of past tense inflection of English verbs. The author thinks that this also can match with the above analysis of the proposed speculation to explain, namely due to the wrong use English past tense does not affect communication misunderstandings in the content, so the subjects in the process of daily communication rarely get timely feedback about the error use of past tense verbs, resulting in the less attention to the type of error. This results in the lack of opportunities for learners to correct mistakes in intermediate language.

It is worth noting that only one grammatical error related to sentence pattern was found in the subjects' oral performance. According to the author, the reason for the low occurrence of this sentence

structure related grammatical error is that complex clauses are rarely used in speech. For speakers, using complex clauses increases the burden of working memory; It also requires more working memory to understand complex clauses. In other words, complex clauses hinder effective communication and lead to potential conversation broken. From the point of view of examination purpose, it is obvious that using complex sentence patterns is also very risky for examinees.

However, the above phenomenon raises a question: can the grammatical knowledge not shown by the subjects in oral expression be regarded as that the subjects have not acquired the grammatical point? For the case analysis of the above three types of errors, it is found that the subjects avoided complex words and sentence patterns which are relatively easy to make mistakes. Can we assume that the subjects have not mastered these grammatical knowledges? The authors are cautious. However, it is questionable to regard the errors not shown by the examinee as the examinee's acquired grammatical knowledge. How to detect the language ability of English learners is not the focus of this topic, so it is not discussed here. However, it should be pointed out that the human-machine dialogue modality of oral test lacks real-time interaction, there is no examiner as the interlocutor to guide the examinee in language expression. This model is still lacking in motivating candidates to demonstrate their mastery of specific second language knowledge.

The fourth type of presentation error discussed in this experiment is logic and content error. Logic and content error refer to that the subject is not perfect in logic and content, or his speech cannot be interpreted because of logic error. The data shows that this type of error is most common in the second half of the human-machine interview (for 2-minute presentations, this type of error tends to occur after the first minute). In the author's opinion, these errors are not so much related to the language ability of the subjects, but to their familiarity with the topic in question. Even native English speakers will make logical errors in content when they make impromptu speeches on unfamiliar topics, or fail to connect the following content logically with what has been said as the presentation progresses. This chapter does not analyze and discuss the specific errors. But worth thinking about the question is, this modality is very strict in time limit, and if the examinee is not familiar with the exam content, he feels "have nothing to say," this may cause logic errors in content. So can it be considered that the candidate's second language ability is not enough? In other words, how can the speaking test in the human-machine dialogue modality technically avoid the failure of the examinees to express their opinions completely due to the time limit and the low judgment on the language competence of the examinees? The above questions are worthy of further consideration and discussion by researchers.

Conclusion

Based on the findings of the experimental, the following conclusions can be drawn: First, the fluency of the subject reflected in the data (including effective speaking frequency and hesitate length in unit time) confirms the scoring distribution of the examiners for the subjects obtained in the previous analysis. In other words, advanced English learners outperform intermediate English learners in speaking fluency. Second, a closer look shows that in addition to their own oral English ability, the subjects' performance is also affected by the topic content. The more familiar the topic is, the better the oral performance of the subjects will be. The influence of content familiarity on the oral fluency of high-level English learners

is significantly higher than that of intermediate English learners. Third, the two types of grammatical errors most ignored by the subjects in oral expression are fixed collocation of phrases and the use of verb tenses. The author believes that these two types of errors are not entirely due to the lack of grammatical knowledge of the subjects, but more to the negative transfer of Chinese to English, which leads to the subjects' failure to form the "expression habit" of English. In addition, the subjects in human-machine dialogue modality cannot get real-time feedback from the communication object, so the subjects are seldom able to notice and self-correct grammatical errors during the expression process. Fourth, the author analyzes and raises research questions about the mistakes made by the subjects in their oral expressions. The human-machine dialogue modality cannot guide the examinees through real-time communication and stimulate the examinees to express the examinee's grasp of relevant second language knowledge which the examiner wants to test. For the sake of "safety score", candidates tend to avoid unfamiliar expressions in favor of the most direct and simple ones. So how can examiners assess candidates' mastery of the second language knowledge that is not involved in their speeches?

References

Abu-Rabia, A. (2001). Bedouin century: Education and development among the Negev tribes in the twentieth century. Berghahn Books.

Alakrash, H. M., Razak, N. A., & Krish, P. (2022). The Application of Digital Platforms in Learning English. International Journal of Information and Education Technology, 12(9).

Alakrash, HM Razak, N, A. (2019). Motivation towards the application of ICT in english language learning among Arab EFL students. Journal of Advanced Research in Dynamical & Control Systems, 11, 1197-1203.

AlSaleem, B. I. (2018). The Effect of Facebook Activities on Enhancing Oral Communication Skills for EFL Learners. International Education Studies, 11(5), 144-153.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford university press.

Huang, I. A., Lu, Y., Wagner, J. P., Quach, C., Donahue, T. R., Tillou, A., ... & Wu, J. (2021). Multi-institutional virtual mock oral examinations for general surgery residents in the era of COVID-19. The American Journal of Surgery, 221(2), 429-430.

Li, K.F., Wang, Y.G. (2018) Artificial Intelligence. Hangzhou(Weekly), 20: 59.

Liu, T., Yuizono, T., Lu, Y., & Wang, Z. (2019, July). Application of human-machine dialogue in foreign language teaching at universities. In *IOP Conference Series: Materials Science and Engineering* (Vol. 573, No. 1, p. 012047). IOP Publishing.

QIAN Cui Jing. 2010. Washback Effect of TEM4 on Oral English Teaching and Learning. Journal of Inner Mongolia University for Nationalities (Social Sciences). 2010-05.

Qian, Y., Ubale, R., Lange, P., Evanini, K., Ramanarayanan, V., & Soong, F. K. (2020). Spoken language understanding of human-machine conversations for language learning applications. *Journal of Signal Processing Systems*, *92*(8), 805-817.

Qian, Z. C., Visser, S., & Chen, Y. V. (2011). Integrating user experience research into industrial design education: The Interaction Design Program at Purdue. In VentureWell. Proceedings of Open, the Annual Conference (p. 1). National Collegiate Inventors & Innovators Alliance.

Ramanarayanan, V. (2020). Design and Development of a Human-Machine Dialog Corpus for the Automated Assessment of Conversational English Proficiency. In *INTERSPEECH* (pp. 419-423).

Ramanarayanan, V., Lange, P. L., Evanini, K., Molloy, H. R., & Suendermann-Oeft, D. (2017). Human and Automated Scoring of Fluency, Pronunciation and Intonation During Human-Machine Spoken Dialog Interactions. In *INTERSPEECH* (pp. 1711-1715).

Razak, H. M., Razak, N. A., & Krish, P. (2022). Enhancing students' digital literacy at EFL classroom: Strategies of teachers and school administrators. *Jurnal Cakrawala Pendidikan*, *41*(3).

Wang, H. (2008). Language policy implementation: A look at teachers' perceptions. Asian EFL Journal, 30(1), 1-38.

Xu Qiang. 2000. Approach to English teaching and assessment [M]. Shanghai: foreign language education press.

Zhang, X. (2013). Foreign language listening anxiety and listening performance: Conceptualizations and causal relationships. System, 41(1), 164-177.

Zheng, Y., & Cheng, L. (2018). How does anxiety influence language performance? From the perspectives of foreign language classroom anxiety and cognitive test anxiety. Language Testing in Asia, 8(1), 1-19.