COMPUTER-DELIVERED ENGLISH LISTENING AND SPEAKING TEST IN ZHONGKAO: TEST-TAKER PERCEPTION, MOTIVATION AND PERFORMANCE

Hengzhi Hu

The National University of Malaysia (UKM), Malaysia, p108937@siswa.ukm.edu.my

Abstract

Against the backdrop that dynamic educational reforms in China have necessitated more flexible assessment of students' proficiency in English as a foreign language, an endeavor has been made to organize computer-delivered English listening and speaking tests (CDELSTs) in the domestic agenda of high-stakes tests, such as *Zhongkao* (also known as the Senior High School Entrance Examination). To understand the complicated factors involved in this innovative assessment mode, the author identified three pertinent variables, namely test-taker perception, motivation and performance, from previous research and examined them quantitatively in a specific Chinese city where the initiative on the organization of CDELSTs has been launched. A random sample of 584 year-nine students who were going to take the CDELST in *Zhongkao* was recruited and completed a mock CDELST and a questionnaire. Descriptive analyses indicated the participants' acceptance of computer-based testing mode, acknowledgement of test importance and considerable effort put into the test, whereas they had reservations about the test validity. Inferential analyses suggested that test performance was correlated with perceived test validity, computer delivery and test-taking effort, with perceived test importance being of no account. Regardless of the discrepancy or compatibility between these findings with previous ones, this research is valuable in the manner that it will inform the test developers and organizers of the voices coming from students as the most pertinent stakeholders and encourage the refinement of CDELST design and administration.

Keywords: computer-delivered test, high-stakes English test, test-taker perception, test-taker motivation, *Zhongkao*

1 INTRODUCTION

In the past two decades, dynamic development and refinement of English language education (ELE) in China have captured considerable attention from the public to Chinese English learners' comprehensive linguistic competencies, and it has been recognized that the mastery of the English language is a personal and national asset that can contribute to China's internationalization push in the context of English being the global lingua franca. Therefore, ELE involved in the Chinese nine-year compulsory education[1] policy landscape has been prioritized more than ever since its legitimation, with 'the earlier the better' belief held by the society (Qi, 2016). Although an opposing and monolingual voice has been lent to the popularization of ELE by reason of the misgiving that it would undermine Chinese students' development of cultural identity and literacy of the first language, it has been generally acknowledged that ELE is an integral part of the basic

[1] Chinese compulsory education system covers six years of primary education with three years of junior secondary education, or five years of primary education with four years of junior secondary education. This difference depends on regional education policies and requirements.

education period[2], without which China cannot develop healthily (Geng & Yuan, 2015).

With a national endeavor initiated in the popularization of ELE to reform curriculum designs, teaching pedagogies and evaluation tasks (Pan, 2014), fostering Chinese students' communicative proficiency has been highlighted given the rapidly growing need for English communication in this Knowledge Age (Amoah & Yeboah, 2021; Liu & Wu, 2015) and the China-specific situation that mute English characterized by the inability to comprehend English listening or speak English well tends to hinder learners' all-round development of language proficiency (Liu et al., 2021). Under this circumstance, considerable attention has been paid to developing learners' English listening and speaking abilities in teaching practices. Aside from this, more efficient assessment of communicative skills has also been called for with an increase in the use of computer-delivered English listening and speaking tests (CDELSTs) in standardized high-stakes tests, such as *Zhongkao*[3] (Liao, 2018). Notwithstanding the criticism about this educational move that may lack assessment validity and reliability (Gao, 2016) or exert a negative washback effect (Huang, 2011), the organization of CDELSTs in domestic examinations has been embraced not only in the most developed cities (e.g., Beijing, Shanghai, Guangzhou) but also in less developed areas. However, with the upsurge of research interest in the standardized CDELSTs organized in those developed areas where innovative educational reforms are normally initiated (Gao, 2016), little has been known from the other contexts wherein this educational trend is followed by local authorities. Meanwhile, the administration of computer-delivered tests (CDTs) is still an emerging educational phenomenon in China, which has not been explored much yet in research with relevant stakeholders (e.g., educators, students, test developers) unaware of the intricate issues interwoven in this assessment mode (Li, 2020; Ren, 2015).

Bearing this brief introduction in mind, the author of this paper wishes to focus on the context of Yantai, a second-tier city in eastern China's Shandong Province, where the CDELST will be included in local *Zhongkao* from the middle of 2022 for all year-nine students[4]. The issues of interest involved in this test can be multifaceted due to the complicated nature of high-stakes tests and CDTs, while test-taker perception, motivation and performance are emphasized in this study as one of the initiatives to add to the growing body of knowledge about innovative assessment methods in China. The research proposals involving these variables are various, whereas the present study primarily intends to explore students' perception of the CDELST in *Zhongkao* from a descriptive perspective as well as to investigate the correlations of the identified variables from an inferential perspective.

2 LITERATURE REVIEW

2.1 CDELST in *Zhongkao*

The implementation of China's basic education reforms since the beginning of the new century has introduced and reinforced the notion of *suzhi jiaoyu* (also known as quality-oriented education[5]), and one of the most constructive policies, which is still playing an essential role, is *China's New National Curriculum Reform* (CNNCR). This policy characterized by the call for the transition from a traditional approach to education to a student-centered one has facilitated the refinement of a comprehensive national curriculum framework that underscores the improvement of educational methods, evaluation measures and knowledge construction, and it has also encouraged the reforms of ELE across all academic levels (Zhou, 2013). In English teaching, a longstanding debate centers around the use of traditional pedagogical approaches (e.g., Grammar-Translation Method) in classrooms, which give prime attention to the teaching and learning of vocabulary and grammar (Liu & Wu, 2015). In a highly examination-oriented context, these approaches have been praised for the practicability of equipping students with sufficient linguistic knowledge to achieve outstanding performance in summative, high-stakes assessment. Nevertheless, they have also been criticized for depriving learners of the opportunities to practice listening and speaking and thus to improve corresponding skills (Pan, 2014). In response to this situation as well as the requirements illustrated in CNNCR, an endeavor has been made at national and regional levels to optimize ELE and promote

[2] According to Zhang et al. (2018, p. 475), China's basic education generally "encompasses pre-school, elementary and secondary education and basic non-formal/informal learning programs".

[3] *Zhongkao*, also known as the Senior High School Entrance Examination, is a standardised test for various school subjects. It is organized annually for year-nine students who intend to continue their studies in senior secondary schools.

[4] Due to the COVID-19 prevention and control policy in the research city, the organization of CDELST in 2022 *Zhongkao* has been decided to be temporarily canceled upon the completion of this paper.

[5] Quality-oriented education aims at the comprehensive development of the qualities (e.g., moral qualities, academic abilities, physical and mental health, personalities) of educated people.

comprehensive language development.

Nationally, for example, the Ministry of Education (MoE) has initiated several rounds of English curriculum reforms in China since the very beginning of this century, and the *English Curriculum Standards for Compulsory Education* (hereafter referred to as the Standards) are the most constructive framework in secondary education and elucidate that comprehensive language use should be promoted both in teaching and assessment (Pan, 2014). Influenced by this policy, various educational initiatives have been launched with special attention to improving and assessing students' English communicative abilities, with Jiangsu Province being the pioneer. At the turn of the century, a paper-based listening test and a face-to-face speaking test started to be organized in Jiangsu Province as a compulsory part of the local *Zhongkao* agenda. However, they were administered separately, which was criticized to be time-consuming and labor-intensive. It was at that time that the proposal of administering the CDELST to secondary school students was formulated, but it was not until 2009 that the test was first implemented, since which it has been a compulsory component of the local *Zhongkao* with the implementation of supporting policies, such as *Outline of the Automated Test of Listening and Spoken English for Junior Middle Schools in Jiangsu Province* (hereafter referred to as the Outline) and *Implementation Measures for the Automated Test of Listening and Spoken English in Junior Middle Schools of Jiangsu Province* (The Editorial Board of the Outline, 2014, 2021). Even though this test has been criticized for a lack of assessment reliability, over-dependence on the automatic scoring system and monotonous design of assessment tasks (Wen, 2016), this pioneering move has been highly commended in academia and inspired the scholars and decision-makers from other parts of China to include or consider including CDELSTs into their regional educational systems (Wang, 2013), and up to now, a number of cities in more than ten provinces have included CDELSTs into local *Zhongkao* (see Fig. 1).

Fig. 1: Chinese cities/provinces where CDELSTs are administered in *Zhongkao*[6]

[6] The highlighted places include Beijing, Tianjin, Chongqing, Liaoning Province, Shandong Province, Jiangsu Province, Anhui Province, Zhejiang Province, Hubei Province, Jiangxi Province, Fujian Province, Guangdong Province, Hunan

Despite the macro development of CDELST at a national level as well as the leadership effect of some pioneering regions on the others, the popularization of CDELSTs is also happening at a micro, provincial level. For example, in Shandong Province, the research context, various cities (see Fig. 2) have included CDELSTs into local *Zhongkao*. The successful implementation of innovative English assessment tasks in these cities has led to the educational reforms in Yantai, and *The Yantai City's Implementation Plan for the Middle School Level Examination in 2020* issued by the Yantai Education Bureau (2020) illuminates that from 2022, the *Zhongkao* English listening test that is originally included in the written test will be replaced by a separate CDELST and that the test score will directly contribute to a candidate's total score in *Zhongkao*.

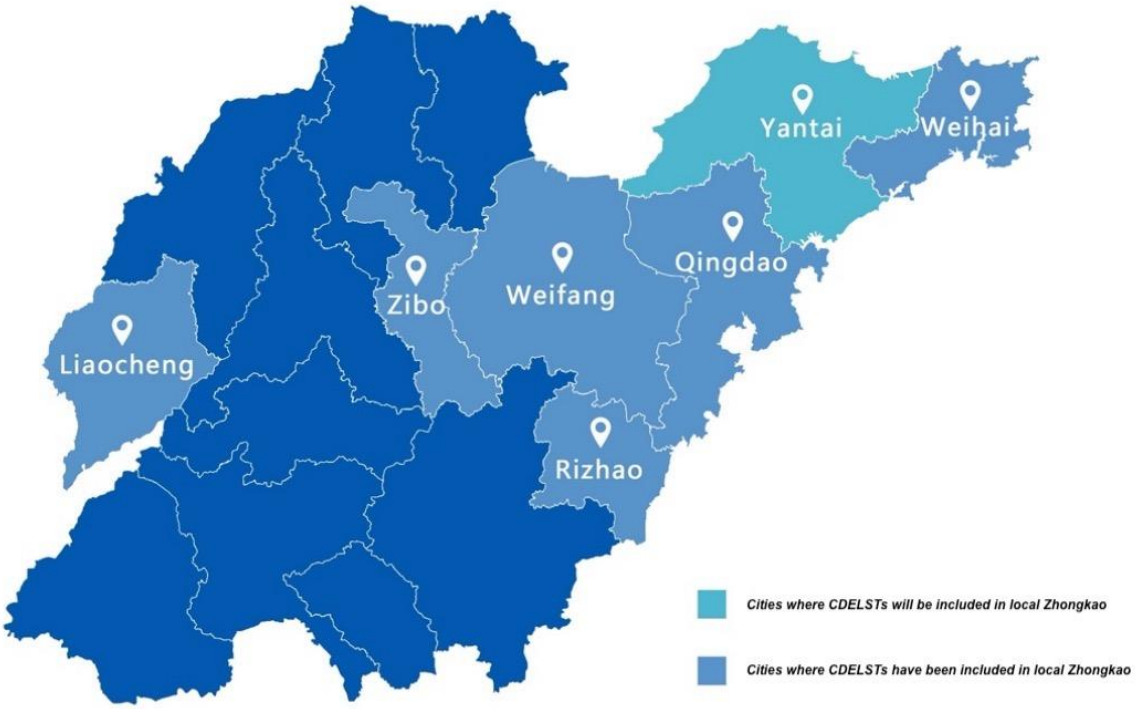

Fig. 2: Cities in Shandong Province of China where CDELSTs have been/will be administered[7]

Similar CDELST structures have been adopted in the *Zhongkao* of different regions. For example, as shown in Table 1, the CDELST to be organized in Yantai consists of three tasks, namely Listening (answering multiple-choice questions), Reading (reading a given text aloud) and Speaking (giving short verbal answers, taking notes and retelling), which weighs 30 marks in total. Regardless of the first listening section which requires objective evaluation, the second section of read-aloud is assessed from the perspectives of Accuracy, Fluency and Completion, and a candidate will lose marks if the following problems are detected from their speech:

- Accuracy: addition[8], omission[9], distortion[10], word stress dislocation, incomplete phrase(s) and/or sentence(s) reading, incorrect pause between thought groups, word stress and/or linking sounds.

Province, Guizhou Province, Yunnan Province, Ningxia Hui Autonomous Region. The author consulted the latest available resources (e.g., newspapers, government announcements) to compile the map. It is acknowledged that some places where CDELSTs are also administered in local *Zhongkao* may not have been included in the map due to a lack of up to date information.

[7] News reports and regional political documents have been consulted to identify the highlighted cities in the figure.

[8] Addition occurs when an extra sound is added.

[9] Omission occurs when a sound is left out. It differs from the omission of words, which occurs when a candidate misses certain words when reading the given text.

[10] Distortion occurs when a sound is not left out but does not sound right. In the original rubrics written in Chinese, it refers to a candidate's inability to read unfamiliar or unknown words.

- Fluency: incorrect use of intonation, rhythm, pause between thought groups, linking sounds, incomplete plosion and/or word stress, dislocation of word stress, distortion, omission of words.
- Completion: incomplete phrase(s) and/or sentence(s) reading, omission of words, distortion, incorrect use of incomplete plosion and pause between thought groups, phonetic elision[11].

The first task (i.e., giving short answers) in the third section is evaluated from the dimensions of Information and Language. The former is associated with the candidate's ability to locate and extract sufficient precise information from the listening materials to give full responses to the questions, while the latter concerns their correct use of syntax, grammar, lexicon and pronunciation. The tasks of notetaking and retelling are done based on the same monologue. Thus, they are both assessed from the criteria of Content, Accuracy and Fluency, and candidates will lose marks if the following issues are detected:

- Content: missing key content information, using the information that is not included in the listening materials, lacking answer extension, using vague expressions.
- Accuracy: sentence fragments, incorrect use of pronunciation, tense, noun-verb agreement and/or word order.
- Fluency: pause and/or incoherence of speech.

Table 1: Details of the CDELST in *Zhongkao*

Section	Format	Requirement	Number of Questions/Task(s)	Score
1: Listening	Multiple-choice questions	listening to dialogues/monologues and choosing the correct answer for each question	10	10
2: Reading	Read-aloud	reading a given text aloud within the time limit	1	5
3: Speaking	Short answers	listening to dialogues and answering the questions verbally	5	7.5
	Notetaking	listening to a monologue and filling the form with necessary information	5	2.5
	Retelling	listening to the monologue used in the note-taking task again and retelling the text within the time limit	1	5

This computer-based assessment is supported by the speech-to-text technology, speech model analysis technology and deep neural network technology, the combination of which is rationalized by the process of manual scoring by experts to establish scoring standards, extracting scoring dimensions, calculating the weight of each dimension, formulating the scoring model and applying it to final assessment (Wei et al., 2019). The CDELST software developed by *Ke Da Xun Fei* (also known as iFLYTEK[12]) has been acknowledged by the MoE as the only feasible and trustworthy application to assist in organizing large-scale online English examinations in China (The Research Group of Computer-based English Listening and Speaking Test, 2012), and it has also been rated highly for its practicability in both education and assessment practices (Luo et al., 2018). Indeed, there were some dissenting voices regarding the reliability of this technology when it was firstly applied in *Zhongkao*, as Jiang's (2010) study revealed that although many students had achieved satisfying scores, their actual level of English proficiency, especially speaking proficiency, was lower than the one required in the Standards. However, the study conducted by Gao (2016) based on multifaceted Rasch analysis has proved the reliability of automatic scoring in CDELSTs for junior

[11] Elision is acceptable in speech, but it may be detected as an omission in the CDELST.

[12] iFLYTEK is a partially state-owned Chinese information technology company featuring voice recognition software and voice-based internet products covering various industries (e.g., education).

secondary school students, indicating that the score earned by a candidate could be a good representation of their actual English abilities.

2.2 Test-taker Perception, Motivation and Performance

The definition of test performance is straightforward, and in a high-stakes test, it simply refers to the band earned by a candidate (Brockmeier et al., 2014). Normally, it plays a vital role in deciding one's further development, as in the case of *Zhongkao*, a student's score directly determines whether they can further their secondary education and even get enrolled into a key senior secondary school (Ryan, 2019). Previous research has mostly focused on the investigation of the factors that can influence test performance in order to optimize education and assessment practices and thus facilitate learning, and the synthesis recorded by Suhaini et al. (2020) has suggested the complicated nature of assessment achievement interwoven with heterogeneous variables. However, amongst these factors, test-taker perception and motivation as two less investigated ones have been recently stressed in different disciplines, such as foreign language (L2) assessment, the understanding of which could facilitate better assessment design and administration as well as more efficient educational practices (Penk et al., 2014; Zhou & Yoshitomi, 2019).

Test-taker perception refers to "students' act of perceiving the assessment in the course under investigation" and has been usually investigated in the field of learning psychology (van de Watering et al., 2008, p. 648). Students' perceived characteristics of an assessment task usually have a profound impact on their approaches to learning and test performance, and they may adjust their learning strategies as per their perceptions when preparing for high-stakes examinations. This view has been verified in various studies. For example, in Razavipour et al.'s (2020) research, the participants' affirmative perceptions of the test content appeared positively correlated with their intensive preparation for the test; in Hoang's (2019) research, the participants' negative perceptions of the test validity interfered with their test performance. Meanwhile, the investigation of test-taker perception can shed light on better assessment design, as Sato and Ikeda (2015), based on their study on students' perceptions of the face validity of a specific assessment task, recommend that test-takers should be informed of the content and ability measured by each test item so that positive washback effects can be exerted.

A relatively new research topic is about test-taker perception of CDTs, and in the field of L2 education, various technology-based tests, especially speaking and listening tests, have been designed and provided in response to the idea that "communicative language ability needs to be conceived in view of the joint role that language and technology play in the process of communication" (Chapelle & Douglas, 2006, p. 108). Previous research findings have presented mixed views toward CDTs, and they particularly center on assessment validity and computer delivery (Zhou & Yoshitomi, 2019). For example, in Amengual-Pizarro and García-Laborda's (2017) survey, students were confident in the face validity of the APTIS speaking test and believed that computer delivery was a user-friendly and valid way to measure speaking proficiency. Similarly, AlAdl's (2020) quasi-experimental study showed that the students taking CDTs had more positive perceptions of high-stakes tests and thus achieved higher scores than those taking traditional paper-based tests (PBTs). In contrast, some differing situations have also been pointed out, such as the one reported by Pathumthong and Jaturapitakkul (2012) who found students' positive perceptions of test delivery and efficiency might not necessarily lead to better test performance and the one presented by Brooks and Swain (2015) who identified that test-takers might take a negative attitude to the Test of English as a Foreign Language Internet-based Test with respect to time constraints and a lack of interaction, as two factors assumed to have a negative influence on their test performance. Such a dichotomy of views is reflected in Zhan and Wan's (2016) study on a CDELST in China for senior secondary school students, who had distinct understandings of the test validity and computer delivery.

Regarding motivation, it generally refers to the process that initiates, directs and continues goal-oriented behaviors, and test-taking motivation maintains the goal-oriented feature and specifically signifies "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). It is believed that students have domain-specific achievement motivation and situation-specific achievement motivation, and test-taking motivation belongs to the latter category because of the situation-specific features of an examination (Penk et al., 2014). Regardless of some pessimistic views that caution researchers against the washback effects that high-stakes tests may have on motivation, such as the one drawn from Dawadi's (2020) latest research that students might have decreased motivation for an important local secondary education examination, it has been generally accepted that test-takers are often motivated in high-stakes tests due to the positive or negative consequences that the tests can cause (Barry & Finney, 2009). This idea has been confirmed in various studies, such as the ones done by Han (2020), Yu and Jie (2013) in the Chinese context, which suggest that

students tend to have high motivation, especially extinct motivation, for high-stakes tests because of the decisive role they play in student academic and vocational development. However, test-takers' levels of motivation can be various in CDTs. On the one hand, the research synthesis done by Ghaderi et al. (2014) reinforces the role CDTs have in enhancing motivation and reducing anxiety. Meanwhile, higher motivation in CDTs usually comes along with better performance, as research has shown that they are positively correlated (AlAdl, 2020). On the other hand, some researchers (Amengual-Pizarro & García-Laborda, 2017; Piaw, 2012), based on empirical studies or comparative analyses of CDTs and PBTs, assert that the test delivery format may be of little effect on test-taking motivation and that CDTs can even aggravate anxiety because of certain personal (e.g., introversion) and external factors (e.g., little expertise in using a computer) and thus influence students' test performance.

Regardless of the heterogeneous views or research findings, some researchers have attempted to draw a link among test-taker performance, perception and motivation. Zhou and Yoshitomi (2019) propose that motivation can be a mediational factor between the other variables by highlighting the situation that negative perception may not necessarily cause low motivation or poor performance; they also spotlight the postulation that the complex nature of motivation based on expectancy-value theory may mediate test-takers' perception and performance, which questions the unproven supposition that negative perception normally causes less effort devoted by students and worse performance made by them. Based on their research on test-takers' perceptions of and motivation for the computer-delivered Test of English for International Communication in Japan, they argue that a candidate's perceived test validity (perception) affects perceived test importance (motivation), which may further contribute to test-taking effort (motivation) along with perceived computer delivery (perception). The test performance is influenced directly by test-taking effort. However, the context-dependent features of this relationship discourage researchers in other contexts from applying it in a general way, just as the case illustrated in the above discussion that previous research on test-taker perception and motivation has generated different and mostly contradictory views, and how they can affect students' test performance still remains to be explored in different educational settings.

With the above as the background and starting point, this study, as one of the first attempts to understand the complex factors interwoven in China's high-stakes CDELSTs, was aimed to provide empirical evidence regarding test-takers' perception, motivation and performance of this test. The participants' general English proficiency was controlled as the moderator variable, because a student's better performance in a test might be related to their higher English ability instead of their positive reactions to the test. The following research questions (RQs) were proposed and addressed in the study:

- RQ1: How do students perceive the validity and computer delivery of the CDELST in *Zhongkao*?
- RQ2: To what extent do students feel motivated for the CDELST in *Zhongkao*?
- RQ3: To what extent are students' perception, motivation and performance related with each other in the CDELST?

3 METHODOLOGY

A cross-sectional survey design was adopted in the study, characterized by the collection of data at one point in time to measure students' perceptions and provide information related to a specific program (Creswell, 2012). With the assistance of the education providers in the research context, a sample of 584 year-nine students who were going to take the *Zhongkao* in 2022 was recruited randomly from a population of over 50,000 students in local secondary schools of Yantai with informed consent obtained from the participants, their guardians and school leaders, and this sample size exceeded the minimum requirement with 95% confidence level and 5% margin of error in social science research (Utts & Heckard, 2015). The sample consisted of 53.6% females (n = 313) and 46.4% males (n = 271) at the age of 14-15, and 70.5% of them (n = 412) came from public schools with the others (n = 172) studying in private ones. All of them had started preparing for the CDELST in *Zhongkao* before the study by taking practice lessons and mock tests.

Three instruments were used in the study. The first one was a mock CDELST on E-Tingshuo[13]. As per the authentic test requirements, three sections of tasks were included (see the above Table 1), the administration of which followed the procedures of pre-test device-checking by invigilators and test takers, responses by candidates and post-test automatic scoring. Although the test included various tasks, only holistic scores were reported by the scoring system and analyzed in the study. The second instrument was the monthly English test organized for each secondary school in the research city. The same test paper

[13] E-Tingshuo is a software program developed by iFLYTEK for young learners' English learning.

developed by the local teaching and research group was used to measure year-nine students' proficiency in vocabulary, grammar, reading and writing, the total score of which was 90. Likewise, only holistic scores were reported in the assessment and were used in this study as indicators of the participants' general English proficiency. The last instrument was a questionnaire consisting of four constructs (i.e., perceived test validity, perceived computer delivery, perceived test importance, test-taking effort) on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The first two constructs measured the test-takers' perceptions of the CDELST, and the other constructs focused on the participants' motivation. The questionnaire items were adapted from the ones developed by Zhou and Yoshitomi (2019) and were translated from English to understandable Chinese by three professional translators who performed and repeated the three-steps framework (i.e., translating, back-translating, cross-checking) (Sperber, 2004). The final Chinese version was pilot-studied, which indicated an acceptable level of reliability (Cronbach alpha = .87) and validity with the factor loading of each item being above .40.

The entire study lasted from September to October 2021. In the beginning, the Chinese questionnaire was distributed to the participants in pencil-and-paper format with the help of their classroom teachers. The response rate was 97.4% (n = 569). Upon the collection of the data, all the questionnaire items were checked for entry accuracy. 23 cases were removed from the analyses due to unfinished items. In October, the English test and the mock CDELST were organized at the participants' schools. The test results were recorded and sent to the researcher by the teachers in charge at each test site. The data were computed into Statistical Package for the Social Sciences 25.0. To answer RQ1 and RQ2, descriptive statistics (e.g., mean, standard deviation, frequency) were computed. To provide an explicit distinction of the students' responses, the choices 'strongly agree' and 'agree' were merged into a general state of 'agree', and the choices 'strongly disagree' and 'disagree' were categorized into overall 'disagree'. Chi-square tests were also run to examine if the frequency amongst the categories was statistically different. To answer RQ3, descriptive statistics and inferential ones generated from Spearman's correlation were computed. If the participants' general English proficiency score was correlated with the CDELST score, semi-partial correlation tests would be run to hold the former constant.

4 RESULTS

4.1 Perception of the CDELST

The descriptive statistics recorded in Table 2 indicated the students had a relatively negative attitude to the validity of the CDELST in *Zhongkao*, with the mean score being 2.35 at the subscale level for the first construct. This was confirmed in the Chi-square tests, the results of which demonstrated that more participants disagreed with the statements about the assessment validity instead of holding a neutral or affirmative view. However, the participants showed slightly differing levels of disagreement with the distinct types of assessment validity, and the mean scores of construct validity displayed by the first two items were slightly higher than those of predictive validity and content validity measured by the last two items respectively. In contrast, the mean score (M = 3.62) of students' perception of the person-to-computer testing mode suggested they had a neutral but moderately positive attitude to it, and the average score of each item of this construct was located between 'neutral' and 'agree' of the scale. The non-significant Chi-square tests suggested a high variation of the responses to each item.

Table 2: Descriptive statistics of test-taker perception

Item	M	SD	Frequency (%)		
			Disagree	Neutral	Agree
Perceived test validity					
1. The abilities measured in the computer-delivered *Zhongkao* Listening Test were key to English communication*.	2.36	.75	64.3	28.6	7.1
2. The abilities measured in the computer-delivered *Zhongkao* Speaking Test were key to English communication*.	2.50	.76	62.5	25.0	12.5
3. Students who receive high scores in the *Zhongkao* CDELST have a high level of English ability*.	2.33	.62	64.6	27.2	8.2

4. The *Zhongkao* CDELST is a good way to test how people use English in daily life*.	2.21	.58	58.8	23.6	17.6
Total	2.35	.67			
Perceived computer delivery					
1. It is natural to talk to a computer during the speaking test.	3.41	.58	28.7	36.6	34.7
2. Taking the *Zhongkao* CDELST is a pleasant experience.	3.57	.51	27.6	29.5	42.9
3. It is okay to ask someone who does not like computers to take the *Zhongkao* CDELST.	3.76	.61	26.8	30.9	42.3
4. It is okay to ask someone who does not have much technical knowledge to take the *Zhongkao* CDELST.	3.75	.65	19.1	36.5	44.4
Total	3.62	.73			

**Chi-square test was statistically significant*

4.2 Motivation for the CDELST

The descriptive statistics presented in Table 3 displayed that the participants attached considerable attention to the CDELST, considering it to be important in their studies with the mean score at the subscale level being 4.26. This was confirmed in the Chi-square tests, indicating that most of the students agreed with the statements about the importance of this English test instead of conveying a negative or neutral point of view. Equally, the participants reported that they put a great deal of effort to this test (*M* = 4.23) on average, with Chi-square tests validating that most students agreed with the statements about test-taking effort and treated the test with great exertion.

Table 3: Descriptive statistics of test-taker motivation

Item	M	SD	Frequency (%)		
			Disagree	Neutral	Agree
Perceived test importance					
1. It is important for me to perform well in the *Zhongkao* CDELST*.	4.21	.98	7.7	15.4	76.9
2. I am concerned about the score I will get in the *Zhongkao* CDELST*.	4.28	.97	5.1	10.3	84.6
3. The *Zhongkao* CDELST is a significant test for me*.	4.36	.90	5.1	5.1	89.8
4. I always want to know how well I can do in the *Zhongkao* CDELST*.	4.18	.94	9.5	11.9	78.6
Total	4.26	.89			
Test-taking effort					
1. I spare no effort to do my best in the *Zhongkao* CDELST*.	4.15	1.04	7.7	11.5	80.8
2. When taking the CDELST, I can persist in completing the	4.35	1.01	8.0	15.4	76.6

tasks[*].					
3. I pay my full attention to the CDELST while completing it[*].	4.19	.92	7.4	11.1	81.5
4. I expend constant hard effort on the CDELST[*].	4.21	1.12	11.8	17.6	70.6
Total	4.23	.98			

[*]*Chi-square test was statistically significant*

4.3 Correlation of Variables

Based on the rule of thumb that 40 < r_s < .70 represents moderate correlation (Guilford, 1956), the data in Table 4 indicated that the students' perception of computer delivery was moderately correlated with the test score, r_s = .60, p < .05. However, no significant correlation could be found between the students' test performance with their perceptions of the test validity, test importance and test-taking effort. As anticipated, the listening and speaking test score was statistically correlated with the participants' general English proficiency (p < .01), suggesting that the latter did confound the relationships between the former with other variables. Thus, semi-partial correlations were computed between the CDELST score and the other variables with the English proficiency score held constant.

Table 4: Descriptive statistics and correlations of the examined variables

Variable	M	SD	1	2	3	4	5	6
1. Perceived test validity	2.35	.67	-	.49	.31	.31	.27	-.06
2. Perceived computer delivery	3.62	.73	.49	-	.44	.40	.60[*]	.33
3. Perceived test importance	4.26	.89	.31	.44	-	.21	.03	-.17
4. Test-taking effort	4.23	.98	.31	.40	.21	-	.16	.01
5. CDELST score	19.23	2.01	.27	.60[*]	.03	.16	-	.85[**]
6. English proficiency	43.20	9.92	-.06	.33	-.17	.01	.85[**]	-

[*]*Correlation was significant at the .05 level (2-tailed)*

[**] *Correlation was significant at the .01 level (2-tailed)*

The inferential statistics of semi-partial correlations presented in Table 5 indicated that the test score was moderately correlated with perceived test validity, r_s = .63, p = .02. Besides, based on the prerequisite that correlation coefficient between .70 to .90 represents a strong correlation (Guilford, 1956), the test score and perceived computer delivery were strongly correlated with each other, r_s = .74, p = .004. Although the correlation coefficient of the test score between perceived test importance was moderate, it was not statistically significant (p > .05). In contrast, the test score was moderately correlated with test-taking effort, r_s = .52, p = .04.

Table 5: Semi-partial correlations between CDELST score with the other variables

Variable	Test Score	*p*
CDELST Score	-	-
Perceived test validity	.63	.02
Perceived computer delivery	.74	.004
Perceived test importance	.47	.11

Test-taking effort	.52	.04

5. DISCUSSION

5.1 RQ1: How do Students Perceive the Validity and Computer Delivery of the CDELST in *Zhongkao*?

The participants' perceptions of the CDELST were explored from the perspectives of test validity and computer delivery. Descriptive and inferential statistics firstly indicated that although students had slightly different attitudes towards the construct, content and predictive validity of the CDELST in *Zhongkao*, their general perceptions were rather negative, suggesting that they were not confident in the test validity. This contradicts previous research findings that students might consider well-designed CDTs as a valid and adequate tool to measure L2 proficiency in speaking, listening or both (Amengual-Pizarro & García-Laborda, 2017; Pathumthong & Jaturapitakkul, 2018) but reflects the ones generated in Yu's (2020) large-scale survey that students might be less satisfied with the content measured in the CDELST included in *Zhongkao*. Due to the positivist nature of this study, the reasons underpinning the participants' doubts about the test validity were not examined. However, Zhan and Wan's (2016) study placed in a similar context suggests that students may believe that the inauthentic design of CDELSTs in China cannot exactly measure the pragmatic competencies required in real-life communication and that there can also be a discrepancy of views on what a test is measuring between test designers and takers. This, to some extent, confirms Jiang's (2010) assumption that some skills required from the candidates in the test to get a high score may not be necessarily used in daily communication, and by using the CDELST organized in Jiangsu Province's *Zhongkao* agenda as an example, she also asserts that the monotonous assessment design may constrain students from showing their real English proficiency (cf. also Wen, 2016). Since the design of the CDELST in the research context is similar to the one adopted in other cities or provinces (e.g., Jiangsu Province), the above-reviewed studies and assumptions make sense in the way that they can be used as one of the possible explanations for the participants' negative perceptions of the test validity.

In contrast, the participants had a slightly positive attitude to the person-to-computer testing mode of the CDELST. Compared with Yu's (2020) survey, this study has presented a different picture in the research city and shown that students could adapt to the innovative computer-based testing mode in high-stakes English tests. The explanation for this can be various, notwithstanding the fact that the participants had prepared for the test and done some practice before the study commenced. For example, Khoshsima et al. (2019) assume that in the current digital age, students have sufficient technical knowledge acquired from both schooling and daily lives to help them to handle CDTs. Besides, although students' preference for CDTs could be various and largely depend on their personal characteristics (e.g., computer familiarity, computer attitudes, computer aversion) (Khoshsima et al., 2019), the author of this paper supposes that the reputation of a high-stakes test and its decisive effects on students' further development could drive them to get familiar with this special testing mode and even force them to accept it by disregarding their divergent characteristics. The rationale of this presumption lies in the idea that the purpose of high-stakes testing to include and exclude people could force test-takers to overcome their deficiencies that are considered unfavorable to their test performance at the cost of developing their strengths or personalities (Carter & Welner, 2013).

It was also found that the perceived test validity was not correlated with the perceived computer delivery, suggesting that even though the participants considered the person-to-computer testing mode to be appropriate in *Zhongkao*, they still harbored reservations about the test validity. Although this reveals a somewhat contradictory situation in the manner that the participants were willing to accept the innovative testing mode but were suspicious about the test validity, a non-negligible, long-establishing voice, though weak, has been reflected. That is, both the public and students have concern about the validity of high-stakes English tests in China (Jin, 2011). However, given that the organization of CDELSTs is still an emerging phenomenon in the research context and even in China, what should be stressed here is Jin's (2014, p. 156) misgivings about "what the test developer and the decision-makers of the test are able and unable to do in their efforts to provide services of a better quality" and whether they have been given sufficient power over the design and the use of the CDTs "to take responsibility for the consequences of the test". This leaves a gap to be filled in future research and requires more negotiation amongst relevant stakeholders to promote the openness and transparency of the design and administration of CDELSTs.

5.2 RQ2: To what Extent Do Students Feel Motivated for the CDELST in *Zhongkao?*

The participants had a high level of motivation for the CDELST in *Zhongkao*, considering it to be important and making considerable effort into it. This fulfills the expectation that students usually have high motivation for high-stakes tests (Barry & Finney, 2009; Penk et al., 2014), especially in the examination-oriented context in China (Han, 2020; Yu & Jie, 2013), and suggests that the affective advantages of CDTs should not be overlooked (Ghaderi et al., 2014). This finding makes sense when embedded in Zhao's (2016, n.p.) research synthesis, which demonstrates that Chinese young English learners' motivation for standardized tests "is shaped by the sociocultural and historical context of the testing system…and vulnerable to adults' attitudes toward test results"[14]. In other words, the examination-oriented educational context in China and the expectations from parents, teachers and even the public of children's superior test performance can motivate students and encourage them to prioritize high-stakes tests and thus make more effort.

Also, no statistical correlation was found between the two types of motivation examined in the study, namely perceived test importance and test-taking effort, contradicting the assumption that the former could contribute to the latter (Zhou & Yoshitomi, 2019) but reproving the finding that no relationship could be drawn between them (Finney et al., 2020). One possible explanation could be Nichols's (2016, p. 929) interpretation of the research finding that high-stakes English tests might not be effective as tools for increasing student motivation or promoting learning, and he supposes that the purpose of these tests "as the cumulation of classes rather than the starting point" and as the benchmark to decide if students are able to matriculate from the current level of English learning to the next normally prevents teachers and students from gaining meaningful knowledge about learning from the assessment and thus adjusting teaching and learning accordingly to motivate learners or meet the deficits identified within students' understanding of the English language. However, the reasons for the situation that perceived test importance is unrelated to test-taking effort can be various, and they are unclear yet because of the positivist nature of this study and a lack of investigation in the others. Probably, the importance of the CDELST in *Zhongkao* as a high-stakes test is simply a fait accompli, and perceived test importance as a variable is not a priority in the Chinese context but should be considered as a general backdrop for research. Nevertheless, what is clear is that Finney et al.'s (2020) suggestion that more attention should be paid to increasing test-taking effort by identifying the pertinent variables other than test importance is constructive, providing researchers with a train of thought in future studies.

5.3 RQ3: To what Extent are Students' Perception, Motivation and Performance Related with Each Other in the CDELST?

A key finding of this study was that the participants' CDELST score was significantly linked with their perceptions of test validity and computer delivery, suggesting that positive perceptions accounted for good test performance. This confirms the previous research findings that these two variables are always positively correlated (Adeyinka & Bashorun, 2012; AlAdl, 2020; Hoang, 2019; Khoshsima et al., 2019) and highlights the necessity that test-takers should be informed of the ability measured by the test items to optimize assessment administration (Sato & Ikeda, 2015) and that they should also be assisted to get suited to the computer-delivered testing mode (Pathumthong & Jaturapitakkul, 2012). Another important finding was that the CDELST score was moderately related to test-taking effort but not with perceived test importance. The former part of this finding is straightforward, as it has been commonly acknowledged in both theories and research that satisfying test performance is usually the return of the motivational effort made by a candidate (Quinn, 2010). However, according to the multidimensional correlation analyses, it was surprising to find that perceived test importance was of little effect in this study. Again, the author believes that the high-stakes features of the CDELST in *Zhongkao* organized in a highly examination-oriented context can be a persuasive explanation here, and these features subordinate Chinese students to the given assessment arrangements. When examinations are used to decide whether students can continue their studies, it goes without saying that most of them will take the importance of examinations, though they may put different degrees of effort or get divergent levels of achievement. Hence, the significance of this CDELST is undeniable, and fortunately, the participants in this study have shown overwhelming approval for that. However, whether it is a valuable variable that can establish relationships between or amongst the others (e.g., test performance) deserves careful consideration, and at least from this study, the answer is negative.

To iterate, Zhou and Yoshitomi (2019, p. 15) propose that "perceived test validity influences test-taking effort

[14] Zhao (2016) also proposes that assessment motivation is subject to age-related characteristics, while this aspect is not considered in this paper given that the CDELST in *Zhongkao* is merely offered to year-nine students of similar ages.

indirectly through the mediation of perceived test importance" and that "unfavorable perception of computer delivery will lead to decreased test-taking effort, which in turn affects test performance". However, this study has presented a simpler framework wherein a candidate's perceived test validity, perceived computer delivery and test-taking effort might account for their CDELST performance (see Fig. 3). These proposed relationships are placed in a highly examination-oriented context in China, which makes it a fait accompli that the importance of high-stakes tests, such as *Zhongkao*, has already been tacitly approved by students. However, it must be noted that this is just an explanatory framework that presents the relationships of variables in the researcher's own context. More complicated casual analyses, as well as large-scale surveys, are needed to polish the superficial thought in this article and prove the reasonability of this framework. Also, the instrument used in this study was a mock CDELST, the administration setting of which might be different from the one in the authentic test. Because of this difference, test-takers' perceptions and levels of motivation may vary greatly when undertaking the test in a much more formal, strict and even stress-provoking context. Since the implementation of CDELSTs is still in the embryonic stage in China, it is manifest that much remains to be explored with respect to its design and administration.

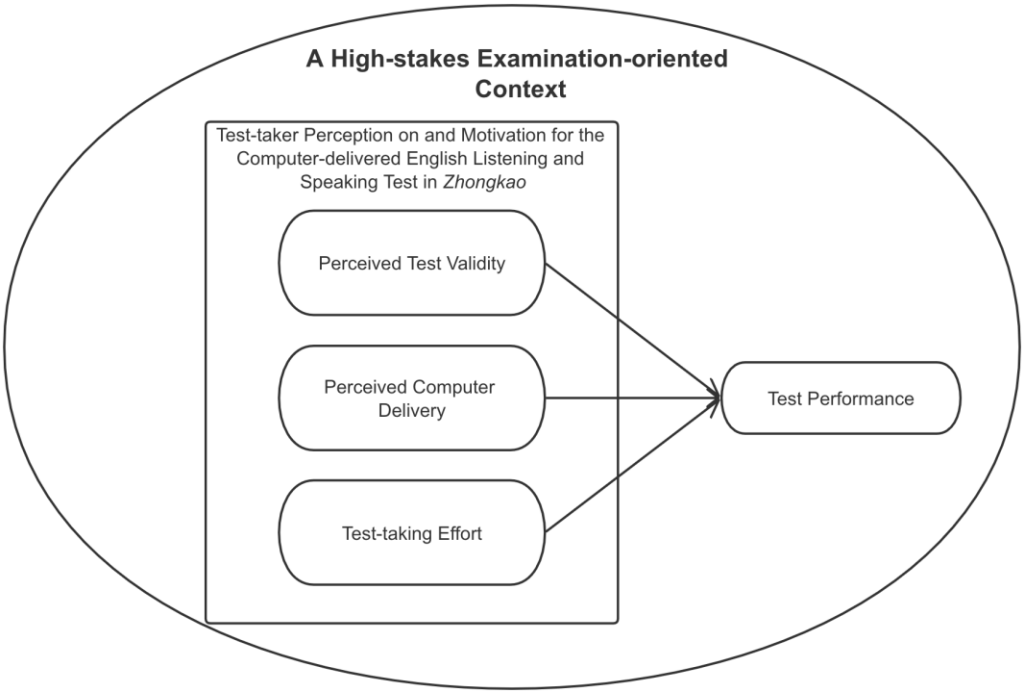

Fig. 3: A proposed framework of the examined variables

6 CONCLUSION

Despite the limitation mentioned above and the drawbacks of quantitative research which could not enable the researcher to understand the participants' experiences and opinions in the studied phenomenon, this study has provided a vital understanding of test-takers' perception and motivation for a high-stakes test, namely the CDELST in *Zhongkao*, in relation with their test performance. To summarize, descriptive statistics demonstrated test-takers' positive attitudes to the computer-based testing mode, approval of the test importance and considerable test-taking effort; inferential statistics suggested that significant correlations could be found between test performance with perceived test validity, perceived computer delivery and test-taking effort. Contradictory to the expectation, however, the participants tended to have a gloomy outlook on the test validity, and the relationship between perceived test importance and test performance was of no account.

Regardless of the discrepancy or comparability between the above findings with previous ones, the insight drawn from this study is crucial as to how assessment practices can be optimized to prepare students for the CDELSTs administered in China's high-stakes testing agenda. Test developers and organizers play an indispensable role in this process, who should not only inform the public, especially test-takers, of the principles of CDELST design and development to bolster the confidence in assessment validity but also listen to test-takers' dissenting voices to refine test design and implementation embedded in a student-

centered philosophy. As one of the initial attempts at understanding the status quo of high-stakes CDTs in Chinese academia, this study also raises the expectations of scholars and researchers, whose future work on the investigation of the multifaceted variables underpinning CDTs will expand the growing body of research and enrich the understanding about innovative assessment practices in China.

REFERENCE LIST

Adeyinka, T., & Bashorun, M. T. (2012). Attitude of undergraduate students towards computer-based test (CBT): A case study of the University of IIorin, Nigeria. *International Journal of Information and Communication Technology Education, 8*(2), 33-45. https://doi.org/10.4018/jicte.2012040103

AlAdl, A. E. (2020). Using electronic tests versus pen and paper tests: The experience of Delta University. *Journal of The Faculty of Education — Mansoura University*, (110), 31-43. https://doi.org/10.21608/maed.2020.147690

Amengual-Pizarro, M., & García-Laborda, J. (2017). Analysing test-takers' views on a computer-based speaking test. *Profile: Issues in Teachers' Professional Development, 19*(1), 23-38. https://doi.org/10.15446/profile.v19n_sup1.68447

Amoah, S., & Yeboah, J. (2021). The speaking difficulties of Chinese EFL learners and their motivation towards speaking the English language. *Journal of Language and Linguistic Studies, 17*(1), 56-69. https://doi.org/10.52462/jlls.4

Barry, C. L., & Finney, S. J. (2009). *Exploring change in test-taking motivation*. Northeastern Educational Research Association.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441-462. https://doi.org/10.1007/BF03173192

Brockmeier, L. L., Green, R. B., Pate, J. L., Tsemunhu, R., & Bochenko, M. J. (2014). Teachers' beliefs about the effects of high stakes testing. *Journal of Education and Human Development, 3*(4), 91-104. https://doi.org/10.15640/jehd.v3n4a9

Brooks, L., & Swain, M. (2015). Students' voices: The challenge of measuring speaking for academic contexts. In B. Spolsky, O. Inbar & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 65–80). Routledge.

Carter, P. L., & Welner, K. G. (2013). *Closing the opportunity gap: What America must do to give every child an even chance*. Oxford University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (5th ed.). Pearson Education.

Dawadi, S. (2020). High-Stakes test impact on student motivation to learn. *European Journal of Educational & Social Sciences, 5*(2), 59-71. https://dergipark.org.tr/en/pub/ejees/issue/57558/781163

Finney, S. J., Satkus, P., & Perkins, B. A. (2020). The effect of perceived test importance and examinee emotions on expended effort during a low-stakes test: A longitudinal panel model. *Educational Assessment, 25*(2), 159-177. https://doi.org/10.1080/10627197.2020.1756254

Gao, M. (2016). A study of constructing and validating an EBB rating scale for a large-scale and low-stakes English oral test of 8th graders. *China Examinations*, (12), 29-38. http://en.cnki.com.cn/Article_en/CJFDTOTAL-KSYJ201612006.htm

Geng, L., & Yuan, A. (2015). Chinese or English Education? A challenge confronted by Chinese government. *US-China Education Review B, 5*(5), 333-341. https://doi.org/10.17265/2161-6248/2015.05.006

Ghaderi, M., Mogholi, M., & Soori, A. (2014). Comparing between computer-based tests and paper-and-pencil based tests. *International Journal of Education & Literacy Studies, 2*(4), 36-38.

http://dx.doi.org/10.7575/aiac.ijels.v.2n.4p.36

Guilford, J. P. (1956). *Fundamental statistics in psychology and education*. McGraw-Hill Book.

Han, Y. (2020). *Parenting styles, academic motivation and performance - academically successful Mainland Chinese students' perspectives* (Publication No. 28121607) [Doctoral dissertation, Miami University]. ProQuest.

Hoang, N. T. H. (2019). Building a validity argument for the use of academic language tests for immigration purposes: Evidence from immigration-seeking test-takers. *Language Education & Assessment, 2*(3), 135-154. https://doi.org/10.29140/lea.v2n3.148

Huang, L. (2011). *Washback on teachers' beliefs and behaviour: Investigating the process*. Foreign Language Teaching and Research Press.

Jiang, X. (2010). A brief analysis of the spoken English test in Jiangsu Province 2009 senior high school entrance examination. *Education Practice and Research*, (5), 16-17. https://doi.org/CNKI:SUN:JYSJ.0.2010-02-011

Jin, Y. (2011). Fundamental concerns in high-stakes language testing: The case of the College English Test. *Journal of Pan-Pacific Association of Applied Linguistics, 15*(2), 71-83. https://files.eric.ed.gov/fulltext/EJ979915.pdf

Jin, Y. (2014). The limits of language tests and language testing: Challenges and opportunities facing the College English Test. In D. Coniam (Ed.), *English language education and assessment: Recent developments in Hong Kong and the Chinese Mainland* (pp. 155-170). Springer.

Khoshsima, H., Toroujeni, S. M. H., Thompson, N., & Ebrahimi, M. R. (2019). Computer-based (CBT) VS. paper-based (PBT) testing: Mode effect, relationship between computer familiarity, attitudes, aversion and mode preference with CBT test scores in an Asian private EFL context. *Teaching English with Technology, 19*(1), 86-101. https://tewtjournal.org/download/11-tewt_19_1_full_issue/

Li, M. (2020). The construction idea of "foreign language teaching and testing integration laboratory" in the new era —— Taking the School of Foreign Languages, Culture and International Exchange of Zhejiang University as an example. *Survey of Education, 9*(1), 47-48. https://doi.org/10.16070/j.cnki.cn45-1388/g4s.2020.01.017

Liao, R. (2018). *Política de planificación familiar y educación secundaria en China. Estudio del caso de la provincia de Guangdong* [Universidad de Salamanca] (Doctoral dissertation). GREDOS.

Liu, D., & Wu, Z. (2015). *English language education in China: Past and present.* The People's Education Press.

Liu, K., Lian, A., & Yodkamlue, B. (2021). Integrating information literacy training in an English-speaking course in the Chinese context. In Atlantis Press (Ed.), *Proceedings of the 17th International Conference of the Asia Association of Computer-Assisted Language Learning (AsiaCALL 2021)* (pp. 29-39). Atlantis Press. https://doi.org/10.2991/assehr.k.210226.004

Luo, D., Xia, L., Zhang, C., & Wang, L. (2018). Automatic scoring of L2 English speech spoken by Chinese middle school students based on deep learning. *Journal of Shenzhen Institute of Information Technology*, (2), 100-104. http://en.cnki.com.cn/Article_en/CJFDTotal-SZXZ201802017.htm

Nichols, J. (2016). Do high-stakes English proficiency tests motivate Taiwanese university students to learn English? *American Journal of Educational Research, 4*(13), 927-930. https://doi.org/10.12691/education-4-13-2.

Pathumthong, P., & Jaturapitakkul, N. (2012). Attitudes of test takers towards the test of English for Thai engineers and technologists (TETET): An innovative computer-based testing. *KMUTT Research and Development Journal, 35*(4), 403-416. https://digital.lib.kmutt.ac.th/journal/loadfile.php?A_ID=501

Pan, L. (2014). *English as a global language in China: Deconstructing the ideological discourses of English in language education*. Springer.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessment in Education, 2*, Article 5(2014). https://doi.org/10.1186/s40536-014-0005-4

Piaw, C. Y. (2012). Replacing paper-based testing with computer-based testing in assessment: Are we doing

wrong? *Procedia — Social and Behavioral Sciences, 64*, 655-664. https://doi.org/10.1016/j.sbspro.2012.11.077

Qi, G. Y. (2016). The importance of English in primary school education in China: Perceptions of students. *Multilingual Education, 6*(1), Article 1 (2016). https://doi.org/10.1186/s13616-016-0026-0

Quinn, G. W. (2010). *Improving test scores in five easy steps: The silver bullet*. Rowman & Littlefield Education.

Razavipour, K., Mansoori, M., & Shooshtari, Z. G. (2020). Test takers' perspectives on an English language test in Iranian higher education: A washback study. *Issues in Educational Research, 30*(3), 1058-1083. https://www.iier.org.au/iier30/razavipour.pdf

Ren, Y. (2015). An efficient way of second language oral assessment — Computer-based testing. *Journal of Juamjusi Education Institute*, (3), 299-300. https://doi.org/CNKI:SUN:JMSJ.0.2015-03-200

Ryan, J. (2019). *Education in China: Philosophy, politics and culture*. John Wiley and Sons.

Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: A potential impact of face validity on student learning. *Language Testing in Asia, 5*, Article 10(2015). https://doi.org/10.1186/s40468-015-0019-z

Sperber, A. D. (2004). Translation and validation of study instruments for cross-cultural research. *Outcome Assessment, 126*(1), 124-128. https://doi.org/10.1053/j.gastro.2003.10.016

Suhaini, M., Ahmad, A., & Harith, S. H. (2020). Factors influencing student achievement: A systematic review. *International Journal of Psychosocial Rehabilitation, 24*(5), 550-560. https://doi.org/10.37200/IJPR/V24I5/PR201720

The Editorial Board of the Outline. (2014). *Outline of the automated test of listening and spoken English for junior middle schools in Jiangsu Province*. Yilin Press.

The Editorial Board of the Outline. (2021). *Outline of the automated test of listening and spoken English for junior middle schools in Jiangsu Province for 2021*. Yilin Press.

The Research Group of Computer-based English Listening and Speaking Test. (2012). *2012 Guangdong Province general Gaokao of English listening and speaking test: High score special training*. University of Science and Technology of China Press.

Utts, J. M., & Heckard, R. F. (2015). *Mind on Statistics*. Cengage Learning.

van de Watering, G., Gijbels, D., Dochy, F., & van der Rijt, J. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *Higher Education, 56*, 645-658. https://doi.org/10.1007/s10734-008-9116-6

Wang, H. (2013). On the teaching of English listening and speaking under the background of "Man-Machine Dialogue". *Reference For Middle School Education*, (10), 86. http://www.cnki.com.cn/Article/CJFDTotal-ZJCA201310068.htm

Wei, S., Wu, K., Zhu, B., & Wang, S. (2019). Speech evaluation technology assists in oral English teaching and evaluation. *AIView*, 73-79. http://www.cnki.com.cn/Article/CJFDTOTAL-DKJS201903010.htm

Wen, P. (2016). *The status quo and problems occurring in the oral English evaluation of junior middle school in Jiangsu Province — Taking Lianyungang as an example* [Master's thesis, Jiangsu Normal University]. https://doi.org/CNKI:CDMD:2.1017.700851

Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing, 11*(4), 324-348. https://doi.org/10.1080/15305058.2011.589018

Yantai Education Bureau. (2020). *The Yantai City's Implementation plan for the middle school level examination in 2020*. Yantai Education Bureau.

Yu, X. (2020). The influence of computerized automatic scoring on junior high school graduates' perception: An empirical analysis. *Examinations Research, 5*(82), 62-67. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&filename=KSYA202005009

Yu, W., & Jie, B. (2013). Washback effect on second language acquisition test: A case study of college English entrance exam. In Atlantis Press. (Ed.), *Proceedings of the 2013 International Academic Workshop on Social Science* (pp. 864-868). Atlantis Press. https://doi.org/10.2991/iaw-sc.2013.196

Zhan, Y., & Wan, Z. H. (2016). Test takers' beliefs and experiences of a high-stakes computer-based English listening and speaking test. *RELC Journal, 47*(3), 363-376. https://doi.org/10.1177/0033688216631174

Zhang, L., Pan, X., & Bai, D. (2018). Overview of basic education in China. In Atlantis Press (Ed.), *Proceedings of the 4th Annual International Conference on Management, Economics and Social Development (ICMESD 2018)* (pp. 474-479). Atlantis Press. https://doi.org/10.2991/icmesd-18.2018.83

Zhao, C. (2016). Motivational effects of standardized language assessment on Chinese young learners. *Cogent Education, 3*(1), Article 1227020. https://doi.org/10.1080/2331186X.2016.1227020

Zhou, Q. (2013). Reflections of English teachers: The quality-oriented education reform in China's middle schools. *Journal of Cambridge Studies, 8*(1), 155-190. https://doi.org/10.17863/CAM.1469

Zhou, Y., & Yoshitomi, A. (2019). Test-taker perception of and test performance on computer-delivered speaking tests: The mediational role of test-taking motivation. *Language Testing in Asia, 9*(1), 1-19. https://doi.org/10.1186/s40468-019-0086-7