A Critical Review of the Computer-based English Listening and Speaking Test

Cecily Ran Liao
University of Macau
Macau

Abstract

The Computer-based English Listening and Speaking Test (CELST) is a large-scale high-stakes norm-referenced test that supports the selection of candidates within Guangdong province for admission to Chinese higher educational institutes. Because of its popularity and important use for the test takers, it is crucial to examine how much the test is useful for its purpose, which determines its consequences and development in the future. The current study aims to put the CELST into the lens of Bachman and Palmer's (1996) usefulness model, drawing on its six features (i.e., reliability, construct validity, authenticity, interactiveness, impact, and practicality) to provide a balanced assessment on thetest usefulness. The analysis suggests that the CELST generally meets the usefulness standard. The combination of its six qualities indicates an acceptable level of usefulness. Based on the findings, some suggestions are also given for the test design and for future research to validate the test.

Keywords: usefulness, reliability, construct validity, computer-based English listening and speaking test

1. Introduction

The computer-based English Listening and Speaking Test (CELST) is a module of English Test in the National College Entrance Examination (NCEE), commonly known as *Gaokao*, which is the largest examination system in China for senior school graduates who intend to study in tertiary educational institutes (Liu, 2010).In fact, the CELST is developed exclusively for test takers in Guangdong province. It was derived from the Computer Oral English Test (COET) which was originally and merely aimed at Guangdong students who planned to study in language-related majors during college education (Zeng, 2010).But in 2011; it was evolved into the CELST for all Guangdong candidates.

The model which the present study draws on is Bachman and Palmer's (1996) test *usefulness*, which is regarded as the most important metric for a test since a test is always developed for its purpose of use (p.17).Test stakeholders concern how much the CELST is useful for its purpose, which determines its consequences and development in the future. A review of relevant literature noticed a certain number of studies measuring the individual test qualities of the CELST, such as test reliability (Cao, 2020; Liu & Wang, 2018), validity evidence (Zeng, 2010; Zhan &Wan, 2016), authenticity (Zhu et al., 2015) and wash back (Huang, 2011; Lin, 2014; Wu, 2016).But an overall test review on its usefulness was barely found. To fill this gap, the current article intends to put the CELST into the lens of Bachman and Palmer's (1996) *usefulness* model, drawing on its six features (i.e., *reliability*, *construct validity*, *authenticity*, *interactiveness*, *impact*, and *practicality*) (p. 18) and integrating previous relevant studies to provide a balanced overall appraisal on its usefulness.

According to what Bachman & Palmer (1996) advised, "the individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test" (p. 18). The notion of this model is depicted in Figure 1.Reliabilityis defined as "a function of the consistency of scores from one set of tests and test tasks to another" (pp. 19-20).A reliable test tool could produce reliable and meaningful information for test participants. The reliability could be affected by measurement errors caused by test characteristics, such as setting, administration, test takers, scoring procedures and test items. According to different sources of error, reliability could be divided into inter-rater reliability (scoring procedures), test-retest reliability (over time), parallel forms reliability (different forms) and internal consistency reliability (across items) (Franzen, 2011).Construct validity alludes to the extent to which the test score interpretation reflects the construct to be measured. It provides evidence to support logical score interpretation. Authenticity refers to "the degree of the characteristics of a given language test task to the features of a TLU task" (p. 23).

A highly authentic test could offer a meaningful generalizable score interpretation, which is also important in construct validity. Interactiveness is characterized as "the extent and type of involvement of the test taker's individual

characteristics in accomplishing a test task, which may include language knowledge, metacognitive strategies, topical knowledge and affective schemata" (p. 25).

It could vary across distinct test characteristics, which indicates its vital link with construct validity as well. The test impact is divided into two levels, a micro level for individuals and a macro level for society and education systems (pp. 29-30). Different from other five qualities which focus on the testuses, practicality denotes the extent to which the available resources could meet the required resources for a test. A practical test is one whose available resources are more than required resources.

Figure 1 The model of Bachman and Palmer's (1996, p. 18) *Usefulness*

Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact+ Practicality

2. Test Overview

2.1 General Description

2.1.1 Test purpose

According to *Guangdong Province University Entrance Examination English Listening and Speaking Test Syllabus* (hereafter CELST syllabus) (The Education Examinations Authority of Guangdong Province (EEAGP), 2018, p.1),CELST is a large-scale high-stakes norm-referenced test developed to examine test-takers' English knowledge, the ability to use English to complete task in a specific context and the competence in integrated use of English knowledge. The score of the test is converted to 20 points figuring in the whole English test score in the NCEE, supporting the selection of candidates within Guangdong province for admission to national higher educational institutes.

2.1.2 Administration

CELST is administered in language laboratories in mid-March every year. Each test taker is allocated with one computer and is required to give answers according to the test audio stimuli. During the test, their responses are recorded and then uploaded to the server for scoring. Results are released in mid-July. While test takers would merely receive an overall score with no test report, test takers could request a re-mark if they have some doubt about their results.

The *Guangdong Province University Entrance Examination English Listening and Speaking Test Notice* (http://eea.gd.gov.cn/ptgk/content/post_3232051.html) (hereafter CELST notice) (EEAGP, 2021) regulated the test clerks' responsibility for administration. As this is a provincial large-scale exam, test clerks from different districts abide by the same test administration regulation. On the test day, test clerks organize test takers to enter the waiting rooms 45 minutes ahead for security checks and identity verification. Test-takers are required to enter preparation rooms 30 minutes ahead to wait and then enter test rooms (i.e., language laboratories) 15 minutes ahead to take a test. In the test room, the test proctors verify the test takers' identity again, distribute blank papers to test-takers for taking notes and instruct them to log in to the test system. After logging in, test takers check the device volume and recorders. The test instructions are presented on the computer screen. Almost all test takers get familiar with the instructions because similar mock tests are available online for test takers to practice. At the end of the test, test takers leave test rooms on a specific route after test proctors ensure all responses are uploaded to the computer server and all note-taking papers are collected.

2.1.3 Test taker

The test is only administered in Guangdong province, so its test takers principally consist of Guangdong secondary school Grade 12 students and others with equivalent educational credentials that pursue the higher education in China (Liu, 2010). The test-takers population reaches over 700,000 on average each year.

2.1.4 Price and developer information

The price for CELST is around 25 RMB, the same as other subjects' tests in NCEE (The exchange rate of US dollar vs. RMB is currently around 1: 6.4). The test developer is *The Education Examinations Authority of Guangdong Province*

(http://eea.gd.gov.cn/). Zhen et al. (2017) described the general test development process. Test designers are usually selected from college educators.

They are divided into three groups respectively responsible for each sub-test design and required to stay in a specific secure location without any communication with the outside world for half a month. In the design phase, they need to select test's suitable materials from hundreds of foreign materials. After the draft test is done, all groups need to exchange the sub-test design and review each other's sub-tests for quality check. Finally, the general leader and all group leaders also need to review all the sub-tests again and settle the final 5-6 parallel versions for the test.

2.2 Testformat and structure

The test normally lasts for 3 days, with 5-6 parallel versions. The number of versions depends on the number of test takers and the test days. Each version lasts for 30 minutes, containing three sub-tests with 5 mins for reading-aloud, 9 mins for role-playing, 7 mins for story-retelling and the rest for instructions, volume checking and responses uploading. The visual scene prompts of each sub-test are primarily in Chinese with English keywords while the audio input and test instructions (audio and visual) are solely in English. No example is attached with task instructions and task instructions are played for once collectively and for another once individually. The structure is listed as follows and a sample test is enclosed in the Appendix.

2.2.1 Part 1 Reading Aloud

In this section, test takers watch a one-minute monologue video clip about geography, nature, history or science normally adapted from BBC, Discovery or National Geographic channel with subtitles of around 100 words, after which the video text is shown on the screen and test takers are given around two minutes for practicing reading. Then test takers listen to the audio again with text shown on the screen. After these, test takers are required to read aloud following the subtitles on the video clip without hearing the audio and to keep the same pace with the video fragments.

2.2.2 Part 2 Role Playing

In this part, test takers are asked to listen to a 2-minute dialogue with around 100 words between 2 roles about real-life topics, such as personal information, daily activities, school life and personal interests. The dialogue is adapted from overseas language teaching materials. Before listening, keywords with Chinese translation and scene information in Chinese language are presented on the screen. After the conversation, they are required to act on one of the roles shown in the dialogue to ask the computer which acts the other role questions and then answer the computer's questions. In the oral interpretation part (i.e. the question-asking part), test takers are given 3 questions in the Chinese language and are required to interpret them into English. Before each interpretation, they have 20 seconds to prepare. After each interpretation, the computer would respond to the questions. In the answering part, test takers are required to answer 5 questions about the content in the conversation and in the oral interpretation part. Before each question, they have 10 seconds to get prepared.

2.2.3 Part 3 Story Retelling

The final section requires test takers to listen to a two-minute narrative monologue adapted from foreign storybooks with around 200 words twice. Similarly, before listening, keywords with Chinese translation and Chinese scene information are presented on the screen to assist test takers to predict what to be heard about. After the audio, they are given one minute to prepare and two minutes to retell the story. Test takers are supposed to use their own wording and cover as much information as possible.

2.3 Test construct

The CELST syllabus (EEAGP, 2018, p.2) states that the test construct includes language knowledge and language use. Language knowledge refers to the English knowledge inclusive of pronunciation, 3500 vocabulary size, grammatical and pragmatic knowledge while language use mainly includes listening and speaking skills. The requirements of specific components may vary across the annually issued syllabus. Below are the descriptions of language use in the test construct of CELST (EEAGP, 2019, pp. 2-3).

- Listening skill requires test takers to comprehend the dialogues and monologues

1) Able to understand the main idea of the input;
2) Able to obtain specific and factual information;
3) Able to make inferences from what is heard;
4) Able to understand speakers' intentions, viewpoints and attitudes.

- Speaking skill requires test takers to appropriately convey meanings and ideas in English

1) Able to inquire and convey factual information, express meanings and ideas;
2) With natural pronunciation and intonation;
3) Able to appropriately use English;
4) Able to employ effective communication strategies.

The responses in reading-aloud and role-playing are graded by Computer-Automated Scoring (CAS) while story-retelling responses are primarily graded by human raters and by CAS just for reference owing to the highly subjective nature of contents (Guan, 2019). The total score of CELST is 60 points, in which 20 points were for reading-aloud, 16 points for role-play, and 24 points for story-retelling. It was converted into 15 points in the past before 2021, but it is now converted into 20 points, counting into the whole score of NCEE English test. The scoring criteria diversify across different sub-tests, which are listed in table 1-3. The reading-aloud lays emphasis on pronunciation and intonation, whereas the role-playing examines candidates' grammatical and vocabulary knowledge as well as their abilities to comprehend the material and capture factual information. Story-retelling integrates the criteria in the previous two sections, additionally testing candidates' fluency and comprehensive abilities to listen and speak (Zhan & Wan, 2016).

Table1 *Reading Aloud Marking Criteria*

Band	Pronunciation and intonation		Speed and content	
	Scores	Descriptors	Scores	Descriptors
A	8-12	Clear and accurate pronunciation; Precise and natural intonation; Coherent and fluent speaking.	6-8	Read as the audio's speed; Content is complete allowing at most three words missing.
B	4-7	Roughly accurate pronunciation; Roughly precise intonation; Roughly coherent speaking.	3-5	Roughly read as the audio's speed but with missing a few words.
C	0-3	Most phonemes are pronounced incorrectly, in inaccurate intonation with incoherent flow.	0-2	Fail to read as the audio's speed with missing a complete sentence or over 10 words.

Table2 *Role Playing Marking Criteria (Score for per question)*

	Language		Information	
	Scores	Descriptors	Scores	Descriptors
Question translation	1.5	Accurate grammar and vocabulary	0.5	Roughlydeliver the informationas required.
	1	Roughly precise grammar and vocabulary; 0.5 point shall be deducted for errors emerging in following facets: Verb tense and voice; Verb transitivity; Subject-predicate agreement.	0	Information is failed to deliver as required.
	0	Serious grammatical and vocabulary errors which cause misunderstanding of the questions, Such as: Incorrect wh-words; Incomplete sentence structure and incoherent meaning.		
Answering questions	0.5	Do not affect the understanding of conveyed information;	1.5	Information is delivered as required.
	0	Result in inability to understand conveyed information;	1	Information is roughly delivered as required although with one or two secondary information missing, or irrelevant information added.

			0	Information is failed to deliver as required.

Table3 *Story Retelling Marking Criteria*

Band	Content		Language		Fluency		Pronunciation	
	Scores	Descriptors	Scores	Descriptors	Scores	Descriptors	Scores	Descriptors
A	8-12	Most of the information is covered.	3-4	Appropriate language.	4-5	Fluent	3	It does not affect the understanding.
B	4-7	Main information is covered.	2	Roughly appropriate language.	2-3	Roughly fluent	2	Though with some errors, it does not affect the understanding.
C	0-3	Partial information is covered.	0-1	With many language errors.	0-1	Not fluent	1	It affects the understanding.

Note. Table 1-3 are all adapted from *Guangdong GaokaoYingyuTingshuoKaoshi[Guangdong Gaokao English Listening and Speaking Test]*, by Baidu, 2021 (https://baike.baidu.com/item/%E5%B9%BF%E4%B8%9C%E9%AB%98%E8%80%83%E8%8B%B1%E8%AF%AD%E5%90%AC%E8%AF%B4%E8%80%83%E8%AF%95/2678957?fr=aladdin).

3. Appraisal of the CELST

Hardly any reliability data could be found either from the test developer's website or any other test-related document. Moreover, through the researcher's enquiry, test developer declined to offer and publicize any reliability statistics with the justification that it is for internal use only, not for the public. Given the importance of the reliability indices to test stakeholders, test developers ought to release the reliability data to validate this is a reliable test. The lack of official reliability statistics indicates the test maybe inadequately reliable. Nevertheless, a handful of studies are observed to examine its reliability. The following analysis looks at the peripheral elements of test characteristics and relevant published researches to estimate various reliability indices.

Liu and Wang (2018) conducted reliability-related research among 455 test takers and 6 scorers, revealing that CELST has good reliability. Zeng (2010) also claimed, 'the CELST reliability was strengthened by its fair administration and rating' (p. 237). With respect to its characteristics, the test development is indeed under rigorous supervision because of its high-stakes nature. While the test would last for three days, the test settings are ensured to be identically equipped under equally rigorous administration. All test rooms around the province are supervised under cameras to avoid any cheating. The test rubric, response and the relationship between input and response also remain the same to ensure fairness. In terms of scoring procedures, the test is marked primarily through CAS, which eliminates rating errors caused by subjective human judgment and raters' fatigue effect. It indicates good inter-rater reliability, which is also verified by Cao's (2020) findings. Guan (2019) also evidenced a good consistency between computerized and human scoring in her study, consolidating the advantage of computerized scoring. In this sense, CELST seemingly holds good reliability.

CELST annually has 5-6 parallel forms to cover the large test-takers population. Zhen et al. (2017) stated that in the test design stage, the way for developers to balance the difficulty among parallel versions is to maintain the same test designers each year and increase the frequency and stringency of the test review. In order to assess the parallel form's reliability and the effectiveness of the artificial difficulty level control, this study employs Coh-Metrix content analysis of the five parallel tests in 2020.Révész & Brunfaut (2013) identified four dimensions as the significant predictors of task difficulty, which are lexical range (i.e., K1 function words), density (i.e., content words), diversity (i.e., concreteness of content words) and causal content (i.e., causal verbs and causal particles). These related indices are selected from Coh-Metrix results for analysis, except for the K1 function words. Table 4 shows the results of these variables calculated by Coh-Metrix. It could be seen that some aspects are highly varied among parallel forms, such as lexical diversity (Mean=75.39, SD=6.44; Mean=56.31, SD=7.99) and causal content (Mean=53.1, SD=14.36; Mean=48.93, SD=16.63)in role-play and story-retelling. Apparently, the difficulty levels among parallel versions are not highly equal, implying that the parallel form reliability was not high, and that the artificial control method Zhen et al. (2017) proposed is not effective. Taken together, the combined reliability of CELST would be considered moderate.

Table 4 *Coh-Metrix results of different variables among the parallel form input materials*

Sub-tests	**Variables**	**Test A**	**Test B**	**Test C**	**Test D**	**Test E**	**Mean**	**SD**
Reading-aloud	DESWLsy (Word length, number of syllables, mean)	1.567	1.426	1.343	1.635	1.529	1.5	0.11
	DESWLsyd (Word length, number of syllables, standard deviation)	0.89	0.659	0.745	1.005	0.841	0.83	0.13
Role-playing	LEXDIVTD (Lexical diversity, MTLD, all words)	78.789	84.637	68.698	70.562	74.311	75.39	6.44
	LEXDIVVD (Lexical diversity, VOCD, all words)	79.52	79.524	90.335	93.886	89.812	86.61	6.66
	WRDFRQc (CELEX word frequency for content words, mean)	2.236	2.419	2.459	2.531	2.227	2.37	0.14
	WRDCNCc (Concreteness for content words, mean)	384.805	387.333	385.916	381.15	398.5	387.54	6.54
	SMCAUSvp (Causal verbs and causal particles incidence)	37.815	45.283	45.627	70.37	66.39	53.1	14.36
Story-retelling	LDMTLD(Lexical diversity, MTLD, all words)	68.817	50.75	59.836	50.25	51.894	56.31	7.99
	LDVOCD (Lexical diversity, VOCD, all words)	67.835	57.337	69.929	52.68	54.529	60.46	7.89
	WRDFRQc (CELEX word frequency for content words, mean)	2.431	2.231	2.528	2.485	2.416	2.42	0.11
	WRDCNCc (Concreteness for content words, mean)	392.25	394.589	390.551	378.299	413.487	393.83	12.66
	SMCAUSvp (Causal verbs and causal particles incidence)	55.838	44.335	49.261	24.876	70.352	48.93	16.63

3.2 Construct validity

The CELST construct validity was validated by two related studies. Zeng (2010) investigated the construct validity of the *Computerized Oral English Test* (*COET*) (CELST was adapted from COET.) through experts' judgments, reporting that the test appears to test what it claims to test. However, Zhan and Wan (2016) conducted research in a converse perspective, collecting test takers' perceptions towards its construct validity, identifying that in some test takers' perspectives, some defined constructs, such as communicative strategic competence, are not as acknowledged in the test as what Zeng's (2010) study claimed.

As could be seen in the above-mentioned test construct, the listening and speaking skills are clearly elaborated respectively into four aspects. Some sub-tests do test what construct intends to test, such as 'able to obtain specific and factual information' and 'natural pronunciation and intonation'. However, some construct components are seemingly not reflected in the tasks.

For instance, asking test takers to translate questions and answer questions would not adequately reflect a construct defined as "able to employ effective communication strategies". Test takers only need to trigger discursive language knowledge and listening strategy to capture factual information, without necessity to invoke communicative strategy. Besides, through looking at the contents in the sample test attached, it could be observed that barely any item reflects the construct component defined as "able to make inferences about what were heard and understand speakers' intentions, viewpoints and attitudes" but only reflects the component defined as "able to get the main idea and obtain specific information", indicating there is a likelihood that not all parallel versions comprehensively reflect the construct. On top of that, the scoring criteria, in an analytic rubric, adequately reflect most construct components, except for the one that reads "able to employ effective communicative strategies". Without adequate reflection in the construct of "able to employ effective communicative strategies", it may not be permitted to make inferences about test takers' ability to speak in the target language. In a nutshell, with some construct components not reflected in the tasks, the CELST's validity could not be regarded as high. Moderate would be more proper.

Zhu et al. (2015) examined the authenticity of the CELST centering on five dimensions -- setting, rubric, input, tasks and context. In the research, the researchers presented a table (as in Table 5) showing the authenticity of input, tasks and context. Besides, they also argued that the test under a computerized setting does not correspond to real-life conversations, which is consistent with Qian's (2009) claim that personal-to-computer is not adequate to evaluate real-life communication skills. Moreover, the instructions are in Chinese and English, which also lowers its authenticity. Zhu et al. (2015) displayed a comprehensive scope of the authenticity level of CELST, but in the researcher's view, the tasks in role-playing and story-retelling, shown in the table, are not that highly authentic.

The role-playing sub-test mainly includes two types of tasks, one of which is translating Chinese questions into English and the other one is using English to answer questions. Rigorously speaking, this section looks more like a translation and listening task than a genuine conversation task. Such viewpoint was also supported by Zhan and Wan's (2016) findings. They collected test takers' attitudes towards CELST design and its wash back on their English learning. Some informants argued that the role-playing part was not as authentic as daily conversations since the tasks could be completed without listening to the speaker and only factual information was required in the tasks without requisite to add any personal idea, which does not correspond to the real conversations. Likewise in the retelling section, test takers are unlikely given opportunities to listen twice and one minute to prepare before rephrasing a speech. But it is understandable that test developers intend to avoid the test being too difficult by considering the unequal teaching quality among different districts. Hence, the tasks in role-playing and story-retelling in Table 5 would be deemed as moderately authentic here. Taking all Zhu et al.'s (2015) five dimensions into account, the CELST maybe more appropriate to be considered as moderately authentic.

Table 5*Authenticity Levels of CELST in Zhu et al., (2015)*

Authenticity	**Input sources**	**Input language**	**Tasks**	**Context**
Read-aloud	High	High	Low	Moderate
Role-Play	Moderate	Moderate	High	High
Retelling	Moderate	Moderate	High	Moderate

3.4Interactiveness

Sparse empirical research on CELST's interactiveness was found, so the following analysis is again based on the researcher's test analysis. Throughout the test, topical knowledge is barely involved in language use. Test takers only need to answer questions using the information obtained from the input material, without the imperative to activate topical knowledge. Personal characteristics are involved especially in terms of general education level and the amount of preparation with a given test. Test takers with higher-level English proficiency and more preparation with mock tests are likely to get more familiar with the test format and perform better. While the test is conducted under computer-person mode which enables some test takers to avoid negative feelings about face-to-face interaction, affective schemata are still somehow engaged. That's because the computers' technical problems emerging in the test could increase students' anxiety levels and, on the other hand, some students could be easily distracted by other participants' voices in the testing condition (Zhan & Wan, 2016).

However, some components are identified as not involved. As seen from the above scoring criteria, grammatical, lexical and phonological knowledge are highly involved in the test whereas pragmatic knowledge is barely noticed. Language function is not engaged either. Just as what was discussed above, task completion merely runs on factual information obtained from the input rather than language functions.

Metacognitive strategy would be considered as somehow involved particularly in story-retelling section where test takers need to formulate a plan to organize what to be included for their response before retelling the story. On this account, CELST would be seen as moderately interactive.

As a high-stakes test module of NCEE, CELST certainly produced a great impact both on individuals and on society and education systems. A few studies shed light on its positive and negative washback effects on teaching and language learning. Huang (2011) and Tan (2018) drew on different districts, examining teachers' attitudes through questionnaires and interviews, and teachers' teaching behaviours through classroom observation. They both revealed CELST's positive and negative washback effects on teaching. More communicative activities, such as role-plays and discussions are inserted into classroom teaching for learners to practice listening and speaking skills, but meanwhile teachers are found to favour test-oriented practices more than real-life tasks. Lin (2014) and Wu (2016) not only collected teachers' attitudes but also students' perspectives on CELST' s washback effects on their language learning, disclosing that most of students were only willing to focus more on test-oriented practices rather than real-life tasks because they lacked time and energy and are under great learning pressure from other subjects. Zhan and Wan (2016) obtained a similar finding that the lightweight and low-level difficulty of CELST's tasks is inadequate to promote learners to put more attention and efforts into this test.

Above published studies focus on the test's impact on individuals other than society and education. The potential consequences on society and education are evaluated as follows. Firstly, with the increasing weight of CELST this year (i.e. As mentioned above, the weight of the score rose from 15 points to 20 points.), educational program would put more attention on learners' listening and speaking skills instead of mere linguistic knowledge and reading and writing skills. Test stakeholders, including schools, teachers, learners and parents, would increasingly value communicative competence, facilitating learners' all-around development in language learning. Besides, NCEE is a crucial, influential and high-stakes test in China, functioning to select students for higher educational institutes and shaping how the society evaluates all stakeholders, i.e., schools, teachers, parents and students (Cheng, 2008). Despite as a module with lightweight, CELST's great impact cannot be overlooked. One single point is also extremely vital for test takers to gap away from other competitors because there are always hundreds of competitors holding the same score. In this way, CELST unquestionably exerts significant consequences on society and education.

On the practicality component, the CELST is relatively more demanding of different types of resources than other tests, which means it is less practical. As was described in the test overview, there are over 700,000 test takers participating in the test each year in the province, which means it demands plenty of test clerks for administration and raters for the human rating (i.e. In story-retelling sub-test). Also, the huge test-takers population requires sufficient space, computers, papers and cameras. Owing to the high-stakes nature of the test and large population, the more stringent procedures also mean more time for test clerks' training and for test taking (i.e.45 minutes' waiting plus 30 minutes' testing). So, it evidently demonstrates that a large amount of human resources, material resources and time are demanded, so CELST is relatively less practical.

4. Overall usefulness

Collectively, the levels of the six qualities in the *usefulness* model are integrated as in Table 6. It displays that most of the individual qualities are evaluated as moderate except for impact as high and practicality as low. In aggregate, the current review would argue that the CELST meets the usefulness standard. The combination of six qualities indicates an acceptable level of usefulness of CELST. Nonetheless, there are some facets in the design of tasks that need to be improved. In specific, 1) test developer needs to publicize reliability data, 2) the test needs to entirely reflect what the construct aims to test, 3) the difficulty levels among parallel forms should be rigorously equivalent and 4) some tasks, such as the role-play, needs to correspond with the authentic language use domain. As it is a consequential test for individuals and society, test developers ought to optimize the test and improve its usefulness to provide a more reliable, meaningful and useful test for society. Furthermore, up to researcher' knowledge, few studies look at some of the individual test qualities, such as reliability, construct validity and interactiveness. Future research can focus on investigating these qualities to validate the test.

Table 6 *Levels of the six qualities in the usefulness model*

Reliability	Construct Validity	Authenticity	Interactiveness	Impact	Practicality
Moderate	Moderate	Moderate	Moderate	High	Low

5. References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Baidu. (2021, March 21). *Guangdong GaokaoYingyuTingshuoKaoshi [Guangdong Gaokao English Listening and Speaking Test]*. *Baiduencyclopedia*. (https://baike.baidu.com/item/%E5%B9%BF%E4%B8%9C%E9%AB%98%E8%80%83%E8%8B%B1%E8%AF%AD%E5%90%AC%E8%AF%B4%E8%80%83%E8%AF%95/2678957?fr=aladdin).

BiliBili Coloum. (2021, February 10). (*Guangdong Gaokao Kouyu Zhenti ABCD Juan Daan in 2020[Guangdong NECC CELT ABCD tests in 2020]*. https://www.bilibili.com/read/cv9760568

Cao, Linlin. (2020). Comparison of Automatic and Expert Teachers' Rating of Computerized English Listening-Speaking Test. *English Language Teaching*, *13*(1), 18-30.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.

The Education Examinations Authority of Guangdong Province (EEAGP). (2018, October 9). 2019 GuangdongshengPutongGaokaoYingyukeTingshuoKaoshiDagang [2019 Guangdong Province College Entrance Examination English Listening and Speaking Test Syllabus (CELST syllabus)]. https://gaokao.chsi.com.cn/news/file.do?method=downFile&id=1733827718&attach=true&hist=false

The Education Examinations Authority of Guangdong Province (EEAGP). (2021, February 26). GuangdongshengPutongGaokaoYingyukeTingshuoKaoshiZhuyiShixiang [Guangdong Province College Entrance Examination English Listening and Speaking Test Notice (CELST notice)]. The Education Examinations Authority of Guangdong Province. http://eea.gd.gov.cn/ptgk/content/post_3232051.html

Franzen M.D. (2011) Test Reliability. In: Kreutzer J.S., DeLuca J., Caplan B. (eds) *Encyclopedia of Clinical Neuropsychology*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-79948-3_2241

Guan, Y.Q. (2019). YingyuTingshuoKaoshiZidongPingfenXiaoduYanjiu [A validity study on automatic scoring of English listening and speaking test in college entrance examination of Guangdong province]. *Shoucang, 10.*

Huang, L.M. (2011). Guangdong 2010 Nian Yingyu Gaokao Gaige Fangan Dui Nongcun Gaozhong Yingyu Tingshuo Ketang Jiaoxue De Fanbo Xiaoying [The backwash effect of the reform plan of Guangdong 2010 college entrance examination on English listening and speaking class teaching in Rural Senior High Schools]. (Master's thesis, Guangzhou University)

Lin, L.J. (2014). Guangdong Sheng GaokaoTingshuoCeshi Dui GaozhongJiaoxue De FanboZuoyong [The backwash effect of CELST in Guangdong Province on senior high school teaching]. (Master's thesis, Guangxi Normal University).

Liu, Q. (2010). The national education examinations authority and its English language tests. *English language assessment and the Chinese learner*, 29-43.

Liu, Y., & Wang, H. (2018). Jiyu DuoyuanGaihua Lilun de YingyuTingshuoKaoshiXinduXiaoduYanjiu [A Study on Reliability and Validity of English Listening and Speaking Test Based on Multivariate Generalization Theory]. *Journal of Changzi University, 35*(3), 104–108.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*(2), 113-125.

Révész, A., &Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*(1), 31-65.

Tan, H. (2018). The washback of CELST of Guangdong on High School English Teachers' Curriculum Consciousness [Master's thesis, Huaibei Normal University].

Wu, P.P. (2016). Guangdong Sheng GaokaoYingyuTingshuoCeshi Dui Shanqu GaozhongYingyuTingshuoJiaoxue De FanboZuoyong – Guangdong Sheng Jiaoyu Keyan "Shiwu" Guihua 2012 Niandu Yanjiu Xiangmu Yanjiu Chengguo [Backwash effect of Computer-based English listening and speaking test (CELST) in Guangdong Province on English listening and speaking teaching in mountain high schools]. *FengkuangYingyu (Jiaoxue version).*

Zeng, Y. (2010). 17 The Computerized Oral English Test of the National Matriculation English Test. *English language assessment and the Chinese learner*, 234.

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.

Zhu, L.F., Luo, Z.L. & Zheng, S.Y. (2015). Guangdong GaokaoYingyuTingshuoKaoshi Zhenshixing Yanjiu [A study on the authenticity of CELST in Guangdong Province]. *Shoucang*, *3*.

Zhen, Z. G., Chen, H.X., Cheng, F.X., Zhang, F., Li, L.R., Ge, B.F., & Jiao, H.W. (2017, March). Jujiao Guangdong Gaokao "Renji Duihua." *English Studies (For Teachers)*, 18–25.

6. Appendix

CELST Version Ain year 2020

Welcome to Computer-based English Listening and Speaking Test. CELST is a module of National Matriculation English Test (Guangdong Version), consisting of three parts. Part A is Reading Aloud. In this part, you are required to watch a video clip and read after the speaker in the video. Part B is Role Play. In this part, you are required to act as a role and complete three communicative tasks: listen to a speaker, ask the speaker three questions and then answer five questions from the computer acting as another role. Part C is Retelling. In this part, you are required to listen to a monologue, and then retell what you have heard in your own words.

PART A Reading Aloud

In this part, you are required to watch a video clip and read after the speaker in the video.

It was the beginning of a totally new way of life. /But there was another very important difference/ between these hunter gatherers and any of their predecessors. /They were settling down. /A huge shift, which would help spark a Europe wide revolution. /Communities settled down and populations expanded. /This could've driven the need to start producing food. /And as farming spread, the landscape was transformed. /Forests were cleared and villages, then towns, then cities would grow, /founded by descendants of the small groups of pioneers /who first entered Europe around 45000 years ago. /Those early Europeans were people just like you and me.

PART B Role Play

In this part, you are required to act as a role and complete three communicative tasks: listen to a speaker, ask the speaker three questions and then answer five questions.

情景介绍(Scene Introduction)

角色：你是 Tom。(Role: You are Tom)

任务(Tasks)：

(1) 与同学 Mary 谈论古扇展览的话题;(Talk about the topic of ancient fan exhibition with classmate Mary;)

(2) 根据谈话内容回答另一同学的提问。(Answer another student's questions according to the conversation.)

生词(New words)：ivory（象牙）

Transcript and answers

M: Hello, Mary. What did you do today?

W: Oh, I've just come back from the City Museum. There's a wonderful exhibition of ancient fans this week.

M: Fans? You mean the fans that cool us down in hot weather?

W: Yeah. This exhibition is about ancient fans of different dynasties in Chinese history. It is part of the Chinese Culture Month organized by the City Museum.

M: Wow! How special!

W: Yes. It was amazing. The exhibition will end on this Saturday. You really should go.

下面请用英语提出三个问题。每个问题有 20 秒的准备时间。当你听见“滴”声时，开始提问。

Now please ask the speaker three questions. You have twenty seconds to prepare the question. When you hear a beep，begin to ask the question.

1. 我从哪里能得到这个展览的票呢？(Where can I get the ticket to the exhibition?)

2. 这些问题是关于什么的呢？(What are the questions about?)

3. 展览上有多少把扇子？(How many fans are there in the exhibition?)

Transcript and answers:

1. Test taker：Where can I get the ticket to the exhibition?

Computer：Well, the museum offers only e-tickets. First, you have to sign up on its website and then watch a short video about the fans in the exhibition. At the end of the video, there are ten questions. If you can answer at least five questions correctly, you will get an e-ticket.

2. Test taker：What are the questions about?

Computer：They are mostly about the history and material of the fans. They are not very difficult. You can always find the answers in the video. The questions are designed to raise the interest of visitors. I think it is a very good way preparing the visitors with basic knowledge of the exhibition.

3. Test taker：How many fans are there in the exhibition?

Computer：There are more than 120 ancient fans in the exhibition. These beautiful fans were made of different materials, including silk, wood, gold, silver and even ivory. These fans are all valuable antiques borrowed from museums all over China. The City Museum has put a lot of money and efforts in this exhibition.

下面用英语回答五个问题。每小题有 10 秒钟的准备时间。当你听见"滴"声时，开始回答。

Now please get ready to answer five questions. You are allowed ten seconds to prepare the answer. When you hear a beep，begin to answer the question.

Transcript and answers:

1. Computer：Who organizes the Chinese Culture Month?

Test taker：The City Museum.

2. Computer：When will the exhibition end?

Test taker：The exhibition will end on this Saturday.

3. Computer：What do you need to do before answering the questions?

Test taker：Watch a short video about the fans in the exhibition.

4. Computer：For what purpose are the questions designed?

Test taker：The questions are designed to raise interest of visitors.

5. Computer：Where are the fans borrowed from?

Test taker：They are borrowed from museums all over China.

PART C Retelling
In this part, you are required to listen to a monologue and then retell it based on what you have heard.

Tom 以为奶奶被邻居骗钱，后来发现邻居只是一个来帮忙的小男孩。

(Tom thought his grandmother was cheated by his neighbor. Afterwards, he found that his neighbor was just a little boy who came to help.)

关键词(Keywords)：grandma（奶奶）young man（年轻人）money（钱）car（汽车）misunderstood（误会）

Transcript:

A Misunderstanding

Tom paid regular visits to his grandma who lived by herself. Last month, he bought her a smart TV and tried to teach her how to use it. But the old lady was a slow learner. Tom felt he was going to lose his patience. One day, grandma told him that he didn't need to teach her anymore, because she had got a helper. Grandma said a young man just became her neighbor. He was quite warm-hearted. Tom didn't believe the young man would help a lonely old lady for no reason. He asked grandma if she had given him some money. " No." Grandma said. But she was going to buy the young man a nice car, because he had been so kind to her. Tom thought the man was a cheater. Knowing that the man would meet grandma at nine o'clock the next morning, Tom managed to get there at nine, too. When he entered the living room, he saw a little boy playing with a red toy car. The boy looked seven years old. Grandma introduced the boy to Tom, calling him "young man". Tom felt ashamed. He had misunderstood everything.

Source. From *Guangdong Gaokao Kouyu Zhenti ABCD Juan Daan in 2020[Guangdong NECC CELT ABCD tests in 2020]*. (2021, February 10). BiliBili Coloum. https://www.bilibili.com/read/cv9760568