Chinese Language Teaching Methodology and Technology

Volume 5 | Issue 1 Article 3

July 2022

Examinees' Affective Preference for Online Speaking Assessment: Synchronous VS Asynchronous

Yuxiao Du
Harvard University

Fangzheng Zhang
Harvard University

Examinees' Affective Preference for Online Speaking Assessment: Synchronous VS Asynchronous

Yuxiao Du (*Harvard University*)
Fangzheng Zhang (*Harvard University*)

ABSTRACT

With technological advancement and the COVID pandemic, online speaking assessment is increasingly used in language teaching. Two modes are developed: online synchronous testing (direct human-to-human interview) and online asynchronous testing (semi-direct human-to-machine interview). Ample literature has explored how each of the two online modes differs from traditional face-to-face speaking assessments. However, few studies have investigated the differences between the two modes, especially in terms of examinees' affective preferences. This study, therefore, compares the extent to which each mode is accepted and favored by test takers and explores why such an affective preference emerges. The participants are 46 college students enrolled in an Elementary Chinese course. They completed a survey that investigates their level of motivation, self-confidence, and anxiety in the two types of online speaking tests. An open-ended question item solicited further explanations from test-takers. Results showed a strong affective preference for synchronous assessment, as manifested by a higher level of motivation and self-confidence and a lower level of anxiety. Possible reasons are discussed based on students' written responses. The study is theoretically significant as it identifies factors on student experience and performance in online speaking assessments. It also provides practical guidance for language teachers in optimizing online oral tests.

Keywords: Speaking, Assessment, Online, Affective, Synchronous, Asynchronous

Speaking assessment has long been a critical part of the evaluation of learners' language proficiency (Weir et al., 2013). Traditionally, speaking tests are direct live interaction interviews, often conducted face-to-face, between human examiners and examinees (Clark, 1979). The examiners both lead the oral interview and evaluate spoken responses. Typical direct speaking tests include Oral Proficiency Interviews (OPI) designed by the American Council on the Teaching of Foreign Languages (ACTFL) and the speaking section of the International English Language Testing System (IELTS). Such direct assessment has been used for over a century and is often acknowledged for its authentic simulation to real-life communicative interactions (Hughes, 2003).

However, with the advancement of technology, online speaking assessments have become increasingly practical and thus popular. Two distinct modes are developed. The first type is online *synchronous* testing that utilizes video-conferencing technology. The tests are administered in similar ways as their face-to-face counterpart, except that the examinees and examiners are physically apart, using a camera and microphone to communicate. This synchronous online testing presents obvious advantages over traditional assessments in test administration because it conveniently simplifies logistical requirements and bridges geographical distances (Nakatsuhara et al., 2017).

The second type of online speaking assessment adopts the *asynchronous* approach, using machines to replace the role of the examiner. In such tests, examinees respond to a pre-determined text, audio, or video prompts (Clark, 1975). The responses are later evaluated by human raters using OPI-style rubrics. Since it combines machine as examiners and human as raters, such asynchronous online testing is also called semi-direct assessment (Quaid & Barrett, 2020). Common examples include Simulated Oral Proficiency Interview (SOPI) and the speaking section of TOEFL iBT before the automated SpeechRater engine was incorporated (ETS, n.d.). This asynchronous testing is accepted by teachers to accommodate a large student body or to rule out the possible unfairness of OPI brought by interviewer variation (Brown, 2003). However, it was also challenged for its lack of authenticity and interaction (Clark, 1986).

Some test designers went a step further and used machine to not only lead the interview but also evaluate the spoken response. This variant of asynchronous online assessment is named automated tests, since they are administered fully automatically through telephone or computer (Bernstein et al., 2010). Like in semi-direct tests, examinees respond to prepared prompts, but the responses are then rated by computer algorithms. A typical example is the Versant test designed by Pearson (2019). However, most automated tests in use now are widely applied standardized assessments because their design remains logistically complicated and challenging. Thus, the feasibility remains low for individual instructors in universities to apply automated tests that are catered to a specific language course. Since this study aims to support microscopic course design on the institutional level, the automated tests are beyond the scope of the current research. Therefore, the asynchronous online assessment in this paper is narrowly defined as the semi-direct human-to-machine interview, precluding the fully automated tests.

Technological advances certainly have increased the practicality of online speaking assessment, but the recent COVID-19 pandemic suddenly makes it one of few, if not the only, choices left for many language instructors. In consequence, it seems appropriate to explore how these online alternatives to traditional face-to-face speaking assessments should be applied in real-world teaching settings.

The current study explores test takers' affective preferences between synchronous and asynchronous online speaking assessments of second language (L2) by looking at how 46 beginning learners of Chinese at a university in the US perceive two modes of online assessments, conducted respectively on *Extempore* asynchronously and on *Zoom* synchronously. To be specific, the paper first reviews relevant literature on online speaking assessments and affective factors in language learning. Then, the paper introduces the methodology, including participants, research context, data collection, and analysis. The paper then proceeds to present the results and discuss the findings of the study. Finally, the authors discussed the limitations of the present study, suggestions for future research, and implications for teaching practices.

Literature Review

Online Speaking Assessments

Constructs of second language speaking competence have always been evolving, and technology has been constantly used to operationalize the assessment of these constructs (Lim, 2018). The academia has been cautious in examining the reliability and validity of new testing modes empowered by technology. In the following, we review important studies on the validation of synchronous and asynchronous online testing as opposed to face-to-face tests.

Research on synchronous online assessment with video-conferencing technology is scarce, but existing findings are generally consistent. As early as 1992, Defense Language Institute Foreign Language Center (DLIFLC) in the US explored the use of satellite-based video technology to assess learners' oral proficiency in Arabic and Russian. This new mode was compared with face-to-face testing, and no significant difference was found in performance, as shown in scores (Clark & Hooshmand, 1992). There did exist an overall preference by test takers for the face-to-face mode, but the finding should be tempered considering that the video-conferencing technology back then was poor in quality and unfamiliar to examinees.

Two recent studies (Craig & Kim, 2010; Kim & Craig, 2012) compared synchronous online testing and face-to-face testing with 40 English as a foreign language learners in Korea. Findings indicated no significant difference in test performance, with similar overall scores and analytic scores (on fluency, functional competence, accuracy, coherence, and interactiveness) across the two testing modes. On the other hand, the quantitative results showed that anxiety level was higher in the face-to-face mode, while interview data indicated a comparable level of comfort and interest between the modes.

Despite their similarities in test administration and assessment validity, the online synchronous assessment still differs from in-person testing in a number of ways. First, in terms of the testing environment, online synchronous testing provides a lot more flexibility for both examiners and test-takers in choosing the exam location, adjusting screen size and parallax (Grayson & Monk, 2003), and accessing external resources. For instance, a test-taker may choose to take the test in a setting that they are familiar and comfortable with (e.g., at home as opposed to an unfamiliar classroom) or can easily reduce the size of the video screen to avoid pressure imposed by examiners "staring" at you. Second, the full reliance on the Internet and audio/video technology may complicate the picture when assessing examinees' performance. For example, Nakatsuhara et al. (2017) reported that a number of test-takers in their study claimed that "sound quality affected their performance, when the sound and transmission were, in fact, both adequate. These differences are essential in understanding that the two modes are *not* identical.

Compared with the scarcity of research on synchronous online testing, asynchronous online testing has been amply researched. A number of studies (e.g., Luoma, 1997; O'Loughlin, 1997; Shohamy, 1994) have jointly provided statistical evidence on the comparability between such tests and traditional face-to-face testing in terms of scores. In that sense, some researchers (e.g., Stansfield & Kenyon, 1992) believe that asynchronous online oral testing can serve as a surrogate to direct face-to-face assessment, although studies looking at the testing process also revealed differences in examinees' output on the discourse level, with language in asynchronous online tests being less sophisticated in general (Luoma, 1997; O'Loughlin, 1997).

On the other hand, research on examinees' perceptions of asynchronous online testing has drawn a more complex picture. Most studies found a preference for face-to-face testing (e.g., Kiddle & Kormos, 2011; Shohamy et al., 1993), the major reason of which lies in the physical proximity between examiners and examinees and which tends to overcome their psychological barrier (Qian, 2009). However, some studies also found that the distance and absence of examiners in asynchronous tests reduce examinees' level of anxiety and are thus favored in that respect (e.g., Song, 2014).

In summary, previous studies have compared both synchronous and asynchronous online speaking testing with traditional face-to-face oral tests. Results are consistent in supporting the validity of both kinds. However, few studies have explored how the two online modes differ

from each other, which forms a research gap that the current study aims to address. Meanwhile, findings on examinees' affective perceptions and attitudes remain mixed. While most test-takers prefer the face-to-face mode, the two online modes do possess certain features that are favored by certain examinees. It remains unclear in regards to how these features will influence examinees' affective perception of the two online modes. To understand the complexity of affective preferences for online testing modes of varying features, we review research findings on affective factors in language learning in the next part.

Affective Factors in Language Learning and Testing

Many linguists have looked into how language learners' emotions and feelings might influence their learning outcomes. These psychological and attitudinal variables are often called *affective factors* (Laine, 1987). The most influential theoretical model of this notion may be the Affective Filter Hypothesis proposed by Krashen (1982), which argued that students' anxiety, low self-confidence, and lack of motivation will function as a mental filter that hinders their acquisition of L2. While the scope of "affective factors" might be broad and may even include external variables such as classroom environment and atmosphere (Guo & Wang, 2013), the major factors discussed by researchers are internal, including *motivation*, *self-confidence*, and *anxiety* (Arnold, 2020; Ni, 2012). This also rationalizes the choice of using these three constructs in this paper to operationalize the measurement of affective factors, which will be explained in detail in the next section on methods.

To begin with, *motivation* is one of the most thoroughly investigated factors in L2 acquisition research (Dörnyei & Ryan, 2015). As the theoretical basis of much motivation research, Gardner's (2001) socio-educational model theorizes L2 learning motivation as instrumental and integrative orientations. Generally speaking, learners with strong and long-lasting motivations of either type are found to have better learning experiences as well as outcomes (Oxford, 1992).

The second important affective factor is *self-confidence*. It is used as the synonym for relevant concepts used in other resources, such as self-efficacy (Zheng et al., 2009), self-esteem, and self-concept (e.g., Rubio, 2018). Students who lack self-confidence are often found to be timid, unwilling to use the target language in class, and thus have trouble acquiring the language (Ni, 2012).

The third major factor, *anxiety*, is described in the language learning context as fear or apprehension that occurs when a learner is supposed to perform in the target language (Guo & Wang, 2013). While the abundant literature on the relationship between anxiety and language performance has produced mixed results (Scovel, 1978), anxiety is still generally recognized as a negative affect that remains unfavored by language learners (Young, 1986). Thus, in this study, anxiety is considered a negative indicator of examinees' preference, while motivation and self-confidence are positive.

So far, researchers have explored rather extensively the roles these affective factors play in language learning. In language testing situations, it is possible that such factors are just as influential, if not amplified. However, for a long period of time, tests had often been evaluated only in terms of their validity and reliability. Affective variables, on the other hand, were often overlooked by test-developers and teachers but might as well influence examinees' performance in the test (Shohamy, 1982). Overlooking affective factors in test design might have led to an inaccurate assessment of language competence. Thus, Shohamy (1982), in a widely quoted

article, called for further research on the extent to which different affective variables influence language assessment.

Since then, anxiety, among all the affective factors in speaking tests, has received wide scholarly attention (e.g., Sayin, 2015; Shi, 2012; Shomoossi & Kassaian, 2009). Results consistently suggest that anxiety is often particularly high in L2 speaking assessment, as this setting potentially elicits both test anxiety and oral communication anxiety at the same time (Horwitz et al., 1986). Causes are found to be multi-faceted, ranging from test takers' individual perception (Sarason, 1984) to test design, including length, techniques, and format (Young, 1991). Thus, it seems reasonable to deduce that the synchronous or asynchronous nature of online speaking assessment might also provoke different levels of anxiety, which also rationalizes the current study. Different from anxiety, other affective factors, including motivation and self-confidence, remain relatively unexplored. The scarce and mixed nature of current findings on affective variables in language tests forms the second research gap that this study aims to address.

Research Questions

To summarize, previous literature provides rich knowledge on the validity of both synchronous and asynchronous modes of online language assessment, often in comparison to traditional face-to-face testing. However, few studies have investigated how they differ from each other. Meanwhile, affective factors in language assessment, except for anxiety, are relatively under-researched. Even for anxiety, results are generally mixed. To address these two research gaps, the current study aims at answering the following two research questions:

RQ1: What are examinees' affective preferences between synchronous and asynchronous online speaking assessments, as manifested by their levels of motivation, self-confidence, and anxiety?

RQ2: Why do such preferences (if any) emerge?

Methods

Participants and Research Context

The participants were 46 college or graduate students in the US enrolled in an Elementary Modern Chinese course that was conducted fully online due to the COVID-19 pandemic. Prior to taking the course, most participants had zero or limited Chinese learning and exposure.

Throughout the academic year, the participants have taken three asynchronous online speaking assessments through the semi-direct platform *Extempore*, and five synchronous online speaking assessments through the video-conferencing platform *Zoom*. The two types of assessments carry the same weight in students' final grades, follow the same evaluation rubrics, and are designed with the same purpose—assessing students' Chinese oral competence in real-life contexts. The types of questions and topics are also consistent in both assessments, which include both specific questions focusing on grammar and open-ended questions that encourage holistic use of their linguistic repertoire. Such features establish the comparability between the two target types of testing.

Extempore is an asynchronous, semi-direct platform that features human-to-machine interactions (Extempore, n.d.). The student could choose any time to complete the test before the

deadline. In each assessment, the student first went through an instruction page that familiarized them with the operations and interface. Then, the student started the assessment by listening to several questions pre-recorded by the teacher. After each question, they had 15 seconds to prepare and 1 minute to speak. Throughout the preparation and speaking, there was a timer on the screen showing the time left. The platform recorded students' answers and submitted them automatically for teachers to review. The teacher then provided audio feedback and numerical evaluation through the platform.

Zoom is a direct video-conferencing platform that allows synchronous "face-to-face" communication among members using cameras and microphones online (Zoom, n.d.). Before attending each test, the students signed up for an 8-minute time slot. As their own time approached, they joined in the meeting room on Zoom. During the test, they interacted by answering questions pre-designed by the teachers. The students were aware that the whole conversation was recorded. The teacher then provided a written feedback form with a numerical evaluation on the course website.

It should be noted that the invigilation methods are not controlled to be the same: examiners in the synchronous exams are allowed to repeat or clarify the question (although this will be taken into consideration in the evaluation process) and to ask follow-up questions while the audio in the asynchronous exams are fixed with the prepared recordings and can be played only once. This design is intentional as the experimenters deem the level of support to be one of the key differences that arise from the distinctive nature of the two online modes. To be specific, synchronous assessment is often designed to examine test takers' communicative competence, which justifies its simulation to real-life conversations where requests for repetition or clarification and follow-up discussions are common and accepted communicative strategies. Meanwhile, asynchronous assessment, at least with current technology, is relatively more rigid in invigilation. Many platforms (including Extempore used in this study) do not yet allow test designers to set the number of times an audio/video prompt can be accessed or offer clarification material that examinees may opt to check during the test. This difference may (or may not) influence test takers' perceived level of support, which will be an interesting topic to explore in the following experiment.

Data Collection

This is a mixed-method study that collects both quantitative and qualitative data to strengthen the results. At the end of the academic year, the students were asked to reflect upon their experience with these two types of assessments by completing a survey based on 5-point Likert Scale (see Appendix for the survey items and variable names). The survey was adapted from Zayed & Al-Ghamdi (2019), and includes two groups (one for synchronous and one for asynchronous) of 15 statements that investigate participants' level of motivation, self-confidence, and anxiety (5 statements for each factor) before and during the two assessments. All items were presented in random sequence. The order in which participants evaluated the two assessments was also randomized. Before taking the survey, the students were informed that the survey was anonymous and was for the benefit of curriculum development, so they should feel free to share any true feelings and thoughts. The responses were documented as quantitative data to reflect students' affective preferences.

The authors have taken measures to ensure the validity and reliability of the adapted survey. To validate the question items, a draft was sent to 5 experienced instructors of the

institution, and they approved its validity with suggestions for minor modifications. Then, the survey was piloted with eight students who had enrolled in the course in previous years, and they agreed that the statements were appropriate in wording and clear in meaning. The reliability of the survey was later supported by measuring the *Coefficient of Cronbach Alpha* for the collected results using SPSS Statistics V27. As shown in Table 1, the intra-reliability coefficients are either above or reaching 0.7. While not all are ideal, they are considered acceptable due to the relatively small sample size (N = 46).

Table 1. *Alpha Reliability Coefficients of the Survey*

Affective Factors	*Cronbach's Alpha*	
	Synchronous Assessment	Asynchronous Assessment
Motivation	0.696	0.689
Self-confidence	0.735	0.835
Anxiety	0.800	0.862

Moreover, one open-ended question item is added at the end of the questionnaire to collect comments and explanations on previous ratings, especially on the items where their ratings differ significantly between the two types of assessment. The responses serve as the data for qualitative analysis. This adds an emic perspective to the study, showing what is viewed as relevant and accountable by the target group members (Dörnyei, 2007).

To answer RQ1, the analysis started with calculating the mean score of the five items for each affective factor. In this way, the participants' average levels of motivation, self-confidence, and anxiety during the two assessments were described in statistics for further comparative analysis. Discussion targets include any preference shown through the statistical difference (higher scores in motivation or self-confidence and lower scores in anxiety). A paired samples t-test was also conducted to ensure that the differences were statistically significant.

To answer RQ2, the researchers conducted bottom-up coding on the qualitative data, looking for common themes mentioned by the participants that might explain their preferences. Together, the quantitative and qualitative data serve as the basis for further discussion and suggestions for the next steps, which will be introduced in the next two sections.

Results and Discussions

RQ1: Which mode do examinees prefer?

The quantitative data show that in all three measurements, students have different perceptions towards the two assessments. To be specific, the mean scores on *motivation* and *self-confidence* were higher for synchronous assessments, while asynchronous assessments earned higher mean scores on *anxiety*. Table 2 presents the detailed descriptive statistics for the survey responses. It should be noted that such differences appear to be especially salient since each and every question item that evaluates motivation or self-confidence received a higher score in synchronous assessment, and all question items evaluating anxiety received a higher score in asynchronous assessment.

To ensure that the differences in mean scores are significant, the authors conducted paired samples t-tests. Results showed that participants' rankings in terms of their motivation ($t = 6.046$, $p < .001$***), self-confidence ($t = 4.521$, $p < .001$***) and anxiety ($t = -3.066$, $p = .004$**) all

differ significantly between the two types of testing. Table 3 summarizes the results of the paired samples t-test.

Table 2. *Descriptive Statistics for the Survey Responses*

Variables	Synchronous		Asynchronous	
	M	*SD*	*M*	*SD*
Group A: Motivation				
A1	3.76	0.98	3.20	1.15
A2	4.46	0.68	3.57	0.99
A3	4.39	0.67	4.11	0.63
A4	4.24	0.73	3.91	0.80
A5	4.41	0.71	3.74	0.92
Group B: Self-confidence				
B1	3.78	0.91	3.24	0.96
B2	4.09	0.69	3.09	1.18
B3	3.76	1.07	3.30	1.08
B4	3.39	0.87	3.00	0.96
B5	3.87	0.77	3.37	1.03
Group C: Anxiety				
C1	3.15	1.12	3.37	1.11
C2	3.61	1.11	3.80	1.15
C3	3.11	1.32	3.93	1.13
C4	3.37	1.36	3.70	1.16
C5	2.59	1.07	3.15	1.10

Table 3. *Results of the Paired Samples t-test*

	Synchronous		Asynchronous				Cohen's
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
Motivation	4.252	.519	3.704	.619	6.046	**< .001*** **	0.891
Self-confidence	3.778	.613	3.200	.819	4.521	**< .001*** **	0.667
Anxiety	3.165	.907	3.591	.918	-3.066	**.004** **	-0.452

Since high motivation and self-confidence are positive indicators of affective preference, and high anxiety is a negative indicator, it seems rather safe to conclude from the quantitative results that examinees demonstrated a clear affective preference for synchronous online speaking assessment as opposed to its asynchronous counterpart. In synchronous settings, they were generally more motivated to excel, more confident in their L2 proficiency, and less anxious about their performance. The quantitative results have highlighted the affective advantages enjoyed by synchronous oral exams, which may offer language teachers a clear-cut reference in thinking about the affective aspects while designing oral assessments.

Meanwhile, it is interesting to notice the relatively large standard deviation for both modes in all three affective indicators as well as each survey item, which range from 0.52 to 1.36. Especially for anxiety, the overall SD gets over 0.90 for both modes and exceeds 1.00 for all the question items. While this is in line with the mixed findings in the existing literature on affective

factors (especially on anxiety), the wide span may also indicate underlying variability in individual perception and understanding of the affective aspects of oral exams. For instance, it seems reasonable to imagine that the absence of a human examiner might reduce anxiety for some students (who are on the introverted side of personality) while being anxiety-inducing for others (who value human interactions and in-person support). Therefore, it may be too early to reach a simple conclusion that synchronous exams are better. The picture might get more complicated when we look into the interview data and explore specific factors that cause such preferences (and seemingly conflicting perceptions) to emerge. The following discussion of RQ2 may thus bring more fine-tuned and thorough messages to instructors in optimizing test design.

RQ2: Why do the preferences emerge?

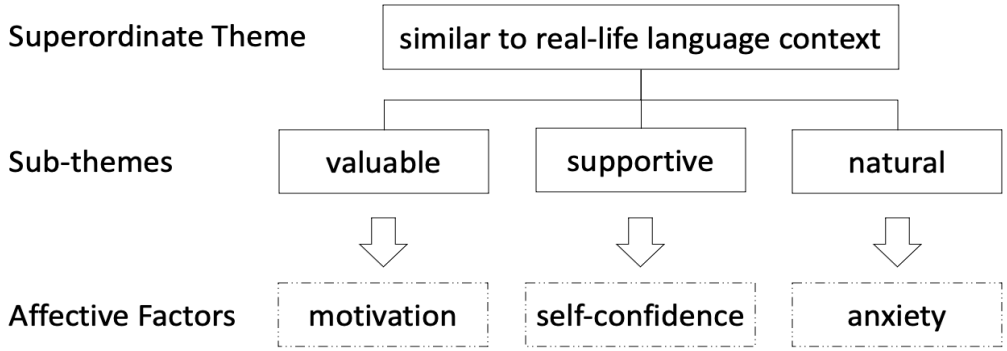
The most common theme that arises from the qualitative responses is the level of *similarity to real-life language context*. A large percentage of students commented that synchronous assessments are more realistic in reflecting real speaking scenarios in our life (see Sample Excerpts No. 1). This "similarity" is co-constructed by multiple features of synchronous assessment, including but not limited to the presence of a real person interacting, as well as the flexibility of test format and content. Details will be further explained below.

Sample Excerpts No. 1:

- I think that the oral exams we did (synchronous) are more realistic as far as future language usage goes.
- Not only do I feel like the oral exams (synchronous) mimic real life more, but they are less nerve-racking and more fun.
- I find that the oral exams (synchronous) are a better reflection of real speaking scenarios and therefore offer a better opportunity to practice my oral Chinese.

Digging further into this superordinate theme, the participants have mentioned different advantages related to this authenticity. Three main sub-themes can be summarized when commenting on the synchronous assessment: *valuable*, *supportive*, and *natural*. Each sub-theme appears to be able to explain correspondingly why students have higher motivation, higher self-confidence, and lower anxiety. See Figure 1 for a visualization of the themes in qualitative responses and how they might explain the affective preference found in quantitative data.

Figure 1. Superordinate and sub-themes on the advantages of synchronous assessment



Sub-theme 1: high motivation for valuable assessment

In the influential Expectancy-Value theory of motivation, Eccles et al. (1983) theorized that students have higher motivation for the type of learning that is perceived to have high *utility value*, i.e., being useful for their future goals. The current responses are in line with the importance of utility value in determining examinees' motivation to prepare and perform well in the assessment. The fact that the synchronous assessment is similar to real-life context helps students perceive it as a good reflection of their Chinese oral proficiency and thus as a helpful tool to prepare themselves for future use of the language.

The high utility value that synchronous assessment enjoys is specially accredited to the fact that live human examiners are able to ask follow-up questions (see Sample Excerpts No. 2), which paves the way for a more well-developed conversation on both sides. As a result, examinees are put under more pressure and motivation to excel. In contrast, the pre-determined nature of interaction in asynchronous assessment seems rather rigid with little flexibility. Such fixed question-and-answer interaction is rare in real life and is thus perceived by test takers to be less valuable and less motivating.

Sample Excerpts No. 2:

- The second (asynchronous) is easy to prepare for but isn't representative of actual conversation. … it probably doesn't provide much value other than being able to remember some set sentences.
- I think the extempore (asynchronous) is a good tool and it helps people when they are just learning how to speak Chinese, but when you do it with the teacher (synchronous) they can ask follow-up questions and it can be more of a conversation and it can be more complex, so I think it is more of an accurate assessment.

Sub-theme 2: high self-confidence in supportive assessment

Self-confidence and performance in competitive situations are often associated with the amount of perceived and received *support* (Rees & Freeman, 2007). Adequate support is found to be crucial in raising participants' self-confidence and thus performance (Fu et al., 2021). In line with such findings, current data foreground the importance of adequate psychological support for examinees that is present in synchronous assessment. In synchronous online oral tests, like conversations in real life, test takers can see live reactions from the interlocutor. Responses common in real-life conversations are also present in the assessment, including but not limited to smiles and frequent nods. When examinees stumble, they might as well be supported by deliberate guidance from the examiners or even just an encouraging look. With such verbal and non-verbal responses as psychological support, examinees may become more confident and thus be more willing to challenge themselves with more complex structures and comprehensive logic (see Sample Excerpts No. 3). On the other hand, the irresponsive computer interface in asynchronous assessment might create an unsupportive environment, leading to less sense of reassurance in the examinees' state of mind.

Sample Excerpts No. 3:

- I strongly prefer the face-to-face oral exams (synchronous). Having a smiling and nodding face to look to while I respond to questions is such a nice atmosphere. I can always ask for something to be repeated, and I feel supported in face-to-face.

- It is much comfortable to talk to a real-person (synchronous) and also I can have real-time feedback.

Sub-theme 3: low anxiety in natural assessment.

Conversations in real-life are often inherently natural, being unaffected and spontaneous, and thus interlocutors are likely to engage in a relaxing interaction. Anxiety is often found to be high in unnatural and unfamiliar contexts (Stansfield & Kenyon, 1992). Our qualitative data show two ways that synchronous assessment outperforms in being natural (see Sample Excerpts No.4). First, the presence of a human examiner (also a responsive interlocutor with emotions) is natural. On the one hand, the examinees are reassured, knowing that they have the options to ask for clarifications or repetitions. On the other hand, the examiners also enjoy the potential to enliven the test with active responses, humor, etc. Second, the testing environment is also familiar and natural. With the development of video-conferencing technology in daily conversations (on Skype, Facetime, etc.), students are already rather used to speaking online with other people. The wide use of Zoom in the US during the pandemic period also familiarizes students with this communicative mode in the educational context. The participants in this study, for instance, have been using Zoom almost every day since the pandemic, whereas Extempore, the asynchronous platform, is only used three times per semester. Therefore, examinees have developed a strong trust in this synchronous communicative platform. To recap, the presence of a natural examiner and familiar testing environment together make synchronous assessment a low-anxiety environment for test-takers.

Sample Excerpts No. 4:

- … Not only do I feel like the oral exams (synchronous) mimic real life more, but they are less nerve-racking and more fun.
- I definitely prefer the oral exams (synchronous) because they are less pressure and you aren't as worried about something going wrong.

Another interesting topic that arises with respect to anxiety is the time limit. Multiple participants reported that the pre-set time limit for each response in asynchronous assessment is rather anxiety-inducing (see Sample Excerpts No.5). Interestingly, in fact, both types of assessment have imposed a time limit on examinees. The major difference is that on the asynchronous Extempore platform, there shows an explicit countdown timer while participants are responding, which is likely the main source of high anxiety. However, for synchronous assessment, there does exist an *implicit* time limit, considering that there is a human interlocutor on the other end waiting for your response. Plus, the whole test for each student has a time limit (8 minutes in this case), so it is a latent mutual presumption that each turn cannot last too long. The fact that participants still reported low anxiety might indicate that this "soft" time limit typical in synchronous assessment is greatly preferred, possibly because it is also present in real life. On the other hand, an explicit timer, which is unfortunately common in many existing asynchronous assessments, may create a great level of anxiety and thus influence students' performance.

Sample Excerpts No. 5:

- … Also, because there isn't a time limit to prepare answers on the oral exam, I can speak more comfortably.
- … it (asynchronous) does impart more anxiety I feel to talk to the computer instead of the person. Particularly in Ba - I felt like the computer assignment

(asynchronous) penalized for not being fast enough whereas the in-person exams are a bit more forgiving for that.

Conclusions, Implications, and Limitations

This study concludes that examinees show a clear affective preference for synchronous over asynchronous online speaking assessment, as shown by their higher motivation, higher self-confidence, and lower anxiety for the former. The fundamental reason might be that the synchronous testing mode is highly similar to a real-life context and thus is perceived to be more valuable, supportive, and natural.

Meanwhile, it is interesting to see that most, if not all, of the advantages of online synchronous assessment mentioned by the participants are also valid for a traditional face-to-face assessment. In the traditional offline oral interview, the interlocutors are also situated in a conversation highly similar to real-life. Examinees are able to ask follow-up questions, see real-time feedback, perceive supportive reactions like smiles and nodding faces, and ask for clarification or repetition. Examiners are free to use humor and soft on imposing a time limit. This high consistency between face-to-face and online synchronous assessment seems to demonstrate that the traits of synchronous oral exams in general, be it online or offline, are generally preferred, compared to their asynchronous counterpart. In other words, it seems somewhat clear that the difference between online and offline assessments in examinees' affect might not be considerable; what really makes the difference is the level of synchronicity between examiners and examinees. Many test-takers nowadays might have already gained enough trust in video-conferencing technology like Zoom to affectively perceive it as a reliable substitute to the traditional face-to-face mode. This preliminary finding, if supported by further empirical evidence in more and wider contexts, may be valuable in setting future research directions in the field of speaking assessment. Knowing that college students have become rather acceptant of video-conferencing technology, scholars could be less worried about transitioning to speaking assessment online and focus more on ways to keep the online assessment synchronous, resembling real-life context.

On the practical level, our findings provide valuable insights for university language teachers on planning assessments, especially in the post-pandemic world where most sectors, including education, start to rethink the integration between in-person and online modes. If conditions permit, instructors could prioritize synchronous tests when speaking assessment needs to be performed online. When designing synchronous assessment, instructors can take measures to enhance its simulation to real life. Possible ways include diversifying tasks (question-and-answer, role play, etc.) and training examiners on establishing a supportive and relaxing testing environment (with smiles, nods, prompt reactions, etc.). Another possible path to optimize the assessment is to control the level of uncertainty perceived by examinees, which is one of the few disadvantages of synchronous assessment mentioned by participants in this study. To achieve that, instructors could provide review guide as a scaffolding measure prior to the test. Moreover, they can also individualize the level of flexibility based on examinees' proficiency and affect. The level of flexibility should be calibrated to help students stay in positive affect and show their true proficiency. Correspondingly, test takers' ability to flexibly engage in the conversation might be incorporated into the rubrics as an important criterion.

Meanwhile, the advantages of asynchronous assessment should not be ignored. Some participants mention that they enjoy the ability to complete the asynchronous test at their

convenience; a few students also point out that they feel less stressed when not talking to a real person. These points, while sporadic, are in line with many previous findings (e.g., Andujar & Cruz-Martínez, 2020; Song, 2014). Thus, the asynchronous assessment might as well function as an alternative or surrogate used in conjunction with its synchronous counterpart.

To better design asynchronous assessment, this study also identifies valuable aspects for instructors and test designers to consider. First, to better simulate real-life context, instructors can prioritize the use of video prompts over audio or text prompts. The videos should be shot with real human interlocutors (ideally someone that the examinees are familiar with). Instructors can also explore virtual world technologies (e.g., secondlife.com) with which examinees become avatars situated in computer-simulated environments. Such technologies, although still new, have received increasing attention from scholars (e.g., Song, 2014) and are found to be useful in creating quasi-authentic conversation contexts. Second, instructors could take remedial measures to lower examinees' anxiety and raise motivation and self-confidence. Several ways can be concluded from the qualitative data: a) use asynchronous technology only in low-stake exams; b) familiarize examinees with the interface in advance; c) offer chances for repetition and clarification; d) countdown time implicitly; e) provide channels for live support.

Nevertheless, due to logistical reasons, the study still bears limitations in sample size and participant diversity. First, all participants are university students aged 18 and above. Second, all participants are students learning elementary-level Chinese. Most of them have no prior experience in learning or speaking Chinese, and some have never been exposed to the language. Thus, the overall content of the course and the topics of oral assessment are all daily and conversational. Third, most participants are born and raised in the US, which means that the participants share a similar cultural background and may only represent US learners.

The above-mentioned homogeneity in participants' age, Chinese level, and cultural background might lead to unrepresentative sampling with respect to participants' age and proficiency, assessment content, and the overall cultural context of the course. However, students of different age groups, with different proficiency levels, or with different cultural backgrounds might have different learning habits and thus may present varying affects towards different modes of oral language testing. Therefore, the scope of the current findings should be limited to university language learners of elementary level in the US. Correspondingly, teachers should take caution while generalizing the results to language courses of more advanced levels, those targeting younger students, or those delivered in non-U.S. cultures / to non-U.S. students. Future studies can further explore examinees' affective attitude towards online oral assessment with younger students, in higher-level courses, or other cultural contexts.

References

Andujar, A., & Cruz-Martínez, M. (2020). Cognitive test anxiety in high-stakes oral examinations: Face-to-face or computer-based? *Language Learning in Higher Education* (Berlin, Germany), *10*(2), 445–467. https://doi.org/10.1515/cercles-2020-2029

Arnold, J. (2020). Affective factors and language learning. In M. Simons, & T. F. Smits (Eds.), *Language education and emotions: Research into emotions and language learners, language teachers and educational processes* (pp. 18-33). Routledge.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1–25. https://doi.org/10.1191/0265532203lt242oa

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones, & B. Spolsky (Eds.), *Testing language proficiency* (pp. 10–28). Arlington, VA: Center for Applied Linguistics.

Clark, J. L. D. (1979). Direct versus semi-direct tests of speaking proficiency. In E. J. Briere, & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 35-49). Washington, DC: TESOL.

Clark, J. L. D. (1986). *Handbook for the development of tape-mediated ACTFL/ILR scale-based tests of speaking proficiency in the less commonly taught languages*. Washington, DC: Centre for Applied Linguistics.

Clark, J. L. D., & Hooshmand, D. (1992). 'Screen-to-screen' testing: An exploratory study of oral proficiency interviewing using video conferencing. *System*, *20*(3), 293–304. https://doi.org/10.1016/0346-251X(92)90041-Z

Craig, D. A., & Kim, J. (2010). Anxiety and performance in video-conferenced and face-to-face oral interviews. *Multimedia Assisted Language Learning*, *13*(3), 9-32. https://doi.org/10.15702/mall.2010.13.3.9

Dörnyei, Z. (2007). *Research methods in applied linguistics*. New York: Oxford University Press.

Dörnyei, Z., & Ryan, S.. (2015). *The psychology of the language learner revisited*. New York: Routledge.

Eccles, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J., & Midgley, C. (1983). Expectancies, values and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75-145). San Francisco, CA: W. H. Freeman.

ETS. (n.d.). *Automated Scoring of Speech*. Retrieved from https://www.ets.org/research/topics/as_nlp/speech/

Extempore. (n.d.). *Why extempore*. Retrieved from https://extemporeapp.com/why-extempore/

Fu, D., Hase, A., Goolamallee, M., Godwin, G., & Freeman, P. (2021). The effects of support (in)adequacy on self-confidence and performance: Two experimental studies. *Sport, Exercise, and Performance Psychology*, *10*(1), 15–26. https://doi.org/10.1037/spy0000206

Gardner, R. C. (2001). Integrative motivation and second language acquisition. In Z. Dörnyei, & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 1-19). Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Grayson, D., & Monk, A. (2003). Are you looking at me? Eye contact and desktop video conferencing. *ACM Transactions on Computer-Human Interaction*, *10*(3), 221–243. https://doi.org/10.1145/937549.937552

Guo, M., & Wang, Y. (2013). Affective factors in oral English teaching and learning. *Higher Education of Social Science*, *5*(3), 57-61. http://dx.doi.org/10.3968/j.hess.1927024020130503.2956

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125-132.

Hughes, A. (2003). *Testing for language teachers* (2nd Ed.). Cambridge, UK: Cambridge University Press.

Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semi-direct test of oral proficiency. *Language Assessment Quarterly*, *8*(4), 342-360. https://doi.org/10.1080/15434303.2011.613503

Kim, J., & Craig, D. A. (2012). Validation of a video-conferenced speaking test. *Computer Assisted Language Learning*, *25*(3), 257-275. https://doi.org/10.1080/09588221.2011.649482

Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press.

Laine, E. (1987). *Affective factors in foreign language learning and teaching: A study of the "filter"* (Jyväskylä cross-language studies; no. 13, 15). Jyväskylä: Department of English, University of Jyväskylä.

Lim, G. S. (2018). Conceptualizing and operationalizing second language speaking assessment: Updating the construct for a new century. *Language Assessment Quarterly*, *15*(3), 215-218. https://doi.org/10.1080/15434303.2018.1482493

Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking: A triangulation study* [Unpublished Licentiate thesis]. University of Jyvaskyla, Jyvaskyla, Finland.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, *14*(1), 1-18. https://doi.org/10.1080/15434303.2016.1263637

Ni, H. (2012). The effects of affective factors in SLA and pedagogical implications. *Theory & Practice in Language Studies*, *2*(7), 1508-1513. https://doi.org/10.4304/tpls.2.7.1508-1513

O'Loughlin, K. J. (1997). *Direct and semi-direct tests of spoken language* [Unpublished doctoral dissertation]. University of Melbourne, Melbourne, Australia.

Oxford, R. (1992). Who are our students? A synthesis of foreign and second language research on individual differences with implications for instructional practice. *TESL Canada Journal*, *9*(2), 30-49. https://doi.org/10.18806/tesl.v9i2.602

Pearson (2019). *Versant English test: Test description and validation summary*. Pearson Knowledge Technologies, Palo Alto, California. Available online at https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-English-Test-Description-Validation-Report.pdf (accessed May 2021).

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. https://doi.org/10.1080/15434300902800059

Quaid, E. D., & Barrett, A. (2020). Toward the future of computer-assisted language testing: Assessing spoken performance through semi-direct tests. In *Recent developments in technology-enhanced and computer-assisted language learning* (pp. 208-235). IGI Global. https://doi.org/10.4018/978-1-7998-1282-1.ch010

Rees, T., & Freeman, P. (2007). The effects of perceived and received support on self-confidence. *Journal of sports sciences*, *25*(9), 1057-1065. https://doi.org/10.1080/02640410600982279

Rubio, F. D. (2018). Self-esteem and self-concept in foreign language learning. In *Multiple perspectives on the self in SLA* (pp. 41-58). Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781783091362-005

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology*, *46*(4), 929. https://doi.org/10.1037/0022-3514.46.4.929

Sayin, B. A. (2015). Exploring anxiety in speaking exams and how it affects students' performance. *International Journal of Education and Social Science*, *2*(12), 112-118.

Scovel, T. (1978). The effect of affect on foreign language learning: A review of the anxiety research. *Language learning*, *28*(1), 129-142. https://doi.org/10.1111/j.1467-1770.1978.tb00309.x

Shi, F. (2012). Exploring students' anxiety in computer-based oral English test. *Journal of Language Teaching and Research*, *3*(3), 446-451. https://doi.org/10.4304/jltr.3.3.446-451

Shohamy, E. (1982). Affective considerations in language testing. *The Modern Language Journal*, *66*(1), 13-17. https://doi.org/10.1111/j.1540-4781.1982.tb01015.x

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, *11*(2), 99-123. https://doi.org/10.1177/026553229401100202

Shohamy, E., Donitsa-Schmidt, S., & Waizer, R. (1993). *The effect of the elicitation mode on the language samples obtained in oral tests*. Paper presented at the 15th Language Testing Research Colloquium, Cambridge, UK.

Shomoossi, N., & Kassaian, Z. (2009). Variation of test anxiety over listening and speaking test performance. *Iranian Journal of Language Studies*, *3*(1), 65-78.

Song, J. (2014). *A study of ESL students' performance and perceptions in face-to-face and virtual-world group oral tests* [Unpublished doctoral thesis]. The University of Texas at Austin, Austin, Texas, US.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*(3), 347-364. https://doi.org/10.1016/0346-251X(92)90045-5

Weir, C. J., Vidakovic, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English examinations, 1913-2012*. Cambridge, UK: Cambridge University Press.

Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals*, *19*(5), 439-445. https://doi.org/10.1111/j.1944-9720.1986.tb01032.x

Young, D. J. (1991). Creating a low-anxiety classroom environment: What does language anxiety research suggest?. *The Modern Language Journal*, *75*(4), 426–437. https://doi.org/10.1111/j.1540-4781.1991.tb05378.x

Zayed, J., & Al-Ghamdi, H. (2019). The relationships among affective factors in learning EFL: A study of the Saudi setting. *English Language Teaching*, *12*(9), 105-121. https://doi.org/10.5539/elt.v12n9p105

Zheng, D., Young, M. F., Brewer, R. A., & Wagner, M. (2009). Attitude and self-efficacy change: English language learning in virtual worlds. *CALICO Journal*, *27*(1), 205-231. https://doi.org/10.11139/cj.27.1.205-231

Zoom (n.d.). *Zoom for education*. Retrieved from https://explore.zoom.us/education

Appendix:

Survey Items and Variable Names

Please rate the following items from 1 to 5 (1-strongly disagree, 2-disagree, 3-neutral, 4-agree, and 5-strongly agree) in two oral assessment settings you've been through:

1) The oral exams we've done through Zoom
2) The speaking assessment we did on Extempore

Question Items	Variable Names
Group A: Motivation	
1. I believe my performance in this assessment is a good reflection of my oral Chinese.	A1
2. I believe preparing for this assessment benefits my oral Chinese significantly.	A2
3. I am determined to do well in this assessment.	A3
4. I think doing well in this assessment is important in helping me get good grades.	A4
5. I think doing well in this assessment is important in helping me become a competent Chinese speaker.	A5
Group B: Self-confidence	
6. I imagine myself as someone who is able to speak Chinese in this assessment.	B1
7. I can imagine myself in the future speaking Chinese fluently in settings similar to this assessment.	B2
8. I think that I am doing my best before and during this assessment.	B3
9. Compared to my classmates, I believe I'm doing relatively well in this assessment.	B4
10. I believe my teacher is going to be satisfied with my performance in this assessment.	B5
Group C: Anxiety	
11. I never feel quite sure of myself when speaking Chinese in this assessment.	C1
12. I am anxious about making mistakes in this assessment.	C2
13. It frightens me when I don't understand what the teacher/audio prompt says in this assessment.	C3
14. I get so nervous when I forget things I know in this assessment.	C4
15. I avoid using complex structures in this assessment.	C5

Open-ended Item:

Please explain any thoughts you would like to share, especially if your answers on the same item differ significantly between the two assessments.

Note: For readers' convenience, the items here are organized into three groups according to their themes. However, on the questionnaire page presented to the participants, the 15 items are all randomized so that the participants were not fully aware of the research aim.