Castledown

OPEN ACCESS

Language Education & Assessment

ISSN 2209-3591

https://www.castledown.com/journals/lea/

Language Education & Assessment, 4 (1), 1–18 (2021)
https://doi.org/10.29140/lea.v4n1.385

Interpretations of spoken utterance fluency in simulated and face-to-face oral proficiency interviews

ETHAN QUAID [a]

ALEX BARRETT [b]

[a] *University of Nottingham Ningbo China, China*
ethan.douglasquaid@outlook.com

[b] *Alasala Colleges, USA*
alex.james.barrett@gmail.com

Abstract

Research examining test taker fluency in simulated and face-to-face oral proficiency interview performances has primarily focused on quantitative spoken utterance fluency data alone, with further qualitative investigation of test taker processing fluency's effect being neglected. This study compared four test takers' spoken utterance and processing fluencies in output retrieved from a computer-based Aptis General speaking test and a purposively developed identical face-to-face direct oral proficiency interview using a counterbalanced research design. Speed, composite, breakdown, and repair utterance fluency measures were analyzed from test performances using Praat speech analysis software and fully-coded transcribed spoken data, with processing fluency qualitative data retrieved through post-test stimulated recall verbal report interviews and questionnaires, and co-constructed, semi-structured interviews. Macro-level quantitative analysis results demonstrated that test takers' spoken utterance fluency was broadly similar between delivery modes. The breakdown measures of filled and unfilled pauses were salient devices responsible for the minimal difference encountered between modes. The test takers' qualitative data reflecting on their performances revealed aspects of processing fluency's effect on test takers' utterance fluency. In turn, processing fluency was influenced by both test takers' cognitive and affective fluencies, through test taking strategy use, test taker characteristics, and the face validity of the test's delivery mode.

Keywords: utterance fluency, processing fluency, speaking test, computer-based assessment

Copyright: *© 2021 Ethan Quaid & Alex Barrett. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*
Data Availability Statement: *All relevant data are within this paper.*

Introduction

Despite today's growing popularity of computer-based simulated oral proficiency interviews (SOPI) being used to assess speaking, there remains insufficient evidence to establish interchangeability with direct face-to-face oral proficiency interviews (OPI), and a need to investigate the equivalence of direct and semi-direct speaking tests from multiple perspectives and collect comprehensive evidence (Alderson, 2004; Weir, 2005; Zhou, 2015). Assessing spoken language with the help of computers may well be a possible source of construct irrelevant variance because delivery mode can be viewed as a facet of task conditions. Chapelle and Douglas (2006) asserted that a primary concern relates to the possibility that test taker performance in SOPIs may not reflect the same abilities as those measured in other forms of assessment.

Further definition of Chapelle and Douglas' (2006) aforementioned broader range of abilities in terms of spoken utterance fluency in output from direct and semi-direct speaking tests has proven somewhat problematic, because attempts to isolate task effect from mode effect and control for interlocutor effect have rarely been wholly successful. Furthermore, the vast majority of previous research investigating test taker spoken utterance fluency in direct and semi-direct speaking tests has been approached purely from a quantitative perspective, with results not being directly related to the test takers themselves, even though they are arguably the most important stakeholders of any speaking proficiency test. However, Zhou (2015) suggested that future research could compare test takers' speech samples, their use of strategies and perceptions of the two modes to complement the concurrent validity evidence presented in previous studies investigating monologic tasked direct and semi-direct speaking tests.

In response to Zhou's (2015) call, and the lack of previous research investigating test taker perception of fluency performance in simulated and face-to-face oral proficiency interviews, this paper expands on a study (Quaid, 2018) that incorporated a mixed method research approach with test takers being given the opportunity to complement their captured spoken utterance fluency data in a SOPI and an OPI through reflection on their performances. The complimentary qualitative data was gained from test takers post-test stimulated recall verbal reports (SRVR) and questionnaires, and co-constructed, semi-structured interviews comparing performances in each delivery mode. Through the addition of this qualitative data, the present study aims to provide insights into test takers' processing fluency in simulated and face-to-face oral proficiency interviews through its respective associations with cognitive and affective fluency.

Terminology

The literature surrounding comparisons of face-to-face and simulated oral proficiency interviews employs several different terms to describe these test modes. References to oral proficiency interviews (OPIs) typically describe a face-to-face delivery of the speaking assessment, involving interaction with a human interlocutor. This is contrasted with computer-based OPIs (CB-OPIs) wherein a computer is used during the elicitation stage of the speaking exam. Furthermore, the term simulated oral proficiency exam (SOPI) has been used to describe more dated forms of speaking tests often entailing the use of a tape recorder or similar device. SOPIs involve relaying speaking prompts to the examinee either on a screen or recording device, with test takers' spoken output being recorded and evaluated by a human examiner at a later time. Similarly, direct and semi-direct OPIs are expressions used to describe the same dynamic, with *direct* meaning face-to-face and *semi-direct* meaning there is a recording device acting as intermediary between the examinee and the examiner. These terms relate to the provision of the exam only, and not the evaluation of the spoken output.

Assessing spoken fluency

Spoken fluency has been notoriously difficult to define because it has several senses when used in speaking testing and assessment that can range from the holistic of spoken proficiency to far narrower definitions that encompass individual features. Therefore, unless fluency is further defined it is often unclear what a speaker or writer is actually referencing (Fulcher, 1996). For example, if fluency measures are not specifically defined for a speaking test rater, a singular indicator may be focused upon with other multiple determinants of fluency in a test taker's speaking test performance left unaccounted for.

De Jong's (2018) review of how fluency is measured by different language exams illuminates the varied understanding of fluency as an aspect of spoken language. In assessment descriptors, according to De Jong, exams typically relate fluency with automaticity, flow, pace, number and placement of pauses or hesitations, presence of repetitions and self-corrections, and duration of speech. Speech is elicited via two main approaches in established speaking assessments: *monologic*, wherein the examinee produces speech individually, and *dialogic*, where an interlocutor is involved. There has been much debate as to which aspects of fluency are most relevant and which method of eliciting speech in assessment contexts is most appropriate.

De Jong *et al.* (2013) maintain that aspects of fluency or disfluency are transferred from a speaker's first language. This illustrates the necessity of understanding which aspects of fluency need to be focused on in second language assessment and which others may be irrelevant. Moreover, many scholars, advocate for task-based oral proficiency exams for eliciting language in fluency assessment, such as one-on-one interviews, which are seen as more closely reflecting real-word interactions (McCarthy, 2010; Sato, 2014; Peltonen, 2017). However, many established second language assessments, such as the Aptis speaking test and Global Test of English Communication, employ monologic tasks. One argument for monologic tasks has been increased reliability, as the interference of an interlocutor's paralinguistic input in an oral proficiency interview does not need to be accounted for.

The many identified potentialities within second language fluency assessment has warranted much attention from researchers who have investigated preferred definitions of fluency, appropriate metrics for assessing it, as well as ideal modes of elicitation. Although fully automated speaking assessments are becoming more popular, this mode of test administration still yields inconsistent results in associated research literature, despite the quantifiable measure of output utterance fluency being focused upon to the detriment of other relevant fluencies that may prove more troublesome to investigate.

Fluencies and measures

Temporal measures of fluency are those that can be measured and quantified. Segalowitz's (2010) framework of inter-related fluencies terms these temporal measures as utterance fluency. Utterance fluency can then be further subdivided into temporal speed, breakdown and repair, and composite measures. The most frequently used method of calculating speed fluency is a speaker's articulation rate. Speakers' articulation rate had long been a relatively unobservable and/or unreliable measure of fluency, yet with the introduction of computer software, such as Praat (Boersma & Weenink, 2013), spoken utterance fluency measures such as articulation rate have become much easier to calculate with a high degree of accuracy.

Breakdown measures include phonation ratio, which is the proportion of time spent speaking minus the time that speakers are silent, and filled and unfilled pause use. These latter two breakdown

measures are particularly important in language testing as they are observable by interlocutors and/or raters and frequently are delineated within speaking test marking rubrics. In contrast, fluency repair can be categorized into word or phrase level repetition, self-correction, and reformulation measures such as the use of false starts at the beginning of utterances; whereas composite fluency measures such as pruned speech rate and mean length of run involve two or more of the aforementioned speed, breakdown, and repair temporal aspects in combination. For instance, if a speaker's speech rate is high, this could be due to either avoidance of unfilled pause use (breakdown measures), and/or a high articulation rate with the production of many syllables (speed fluency measure). Moreover, pruned speech rate could be seen as combining all three aspects of fluency because the pruning speech frequently involves removing syllables that are used to repair an utterance (fluency repair measure).

Unlike the quantifiable aspects of speech production, processing fluency is a metacognitive experience which describes the degree of ease corresponding to the mental processing of information (Graf *et al.*, 2018), which in the present framework refers to speech production. As a metacognitive function, this perceived ease in the processing of information related to speech production is associated with other types of fluency. For example, in response to a speaking test task, an examinee will employ utterance and cognitive fluencies (Segalowitz, 2010), and will also exhibit affective fluency (Jaud & Melynk, 2010) in response to exam conditions. Of the former two fluencies, cognitive fluency can be defined as the speaker's ability to efficiently coordinate many different, interacting cognitive strategies (Segalowitz, 2010) such as planning the utterance, retrieving the appropriate lexis, and grammaticizing. Under exam conditions, cognitive fluency can also include the use of test taking strategies needed to perform well on a test task. Utterance fluency, mentioned in the previous paragraph, relates to the measurable aspects of speech, such as articulation, pausing, and repair measures. Together, cognitive and utterance fluency do much to influence the perceived fluency of an individual by an interlocutor examiner. Affective fluency is another factor which influences the test taker. Affective fluency is defined by Jaud and Melynk (2010) as "the enjoyment of using mental/cognitive resources" (p. 2) towards a task. In exam conditions, this could be affected by test-taker attitude toward the topic of the exam task, or the face validity of the exam. This can in turn generate effects such as anxiety, motivation, and can influence self-confidence (Reber *et al.*, 2004). Affective fluency may not directly affect the perceived fluency of an examinee in a speaking test, but likely has a reciprocal influence on processing fluency, and therefore may indirectly influence other fluencies. A visual representation of how these fluencies and measures thereof could be observed in this present study is provided in Figure 1.

Within the domain of testing speaking, and through alternate test delivery modes in particular, no previous research has investigated processing, cognitive, or affective fluency, and has instead more often than not focused on face validity (Qian, 2009). Moreover, due to the preference for large-scale quantitative method studies within the language testing and assessment research community that satisfy test developers and administering institutions, equally important stakeholders such as test preparation teachers and test takers themselves are often left unaccounted for, which only serves to widen the chasm between language testing research and test preparation classroom practices.

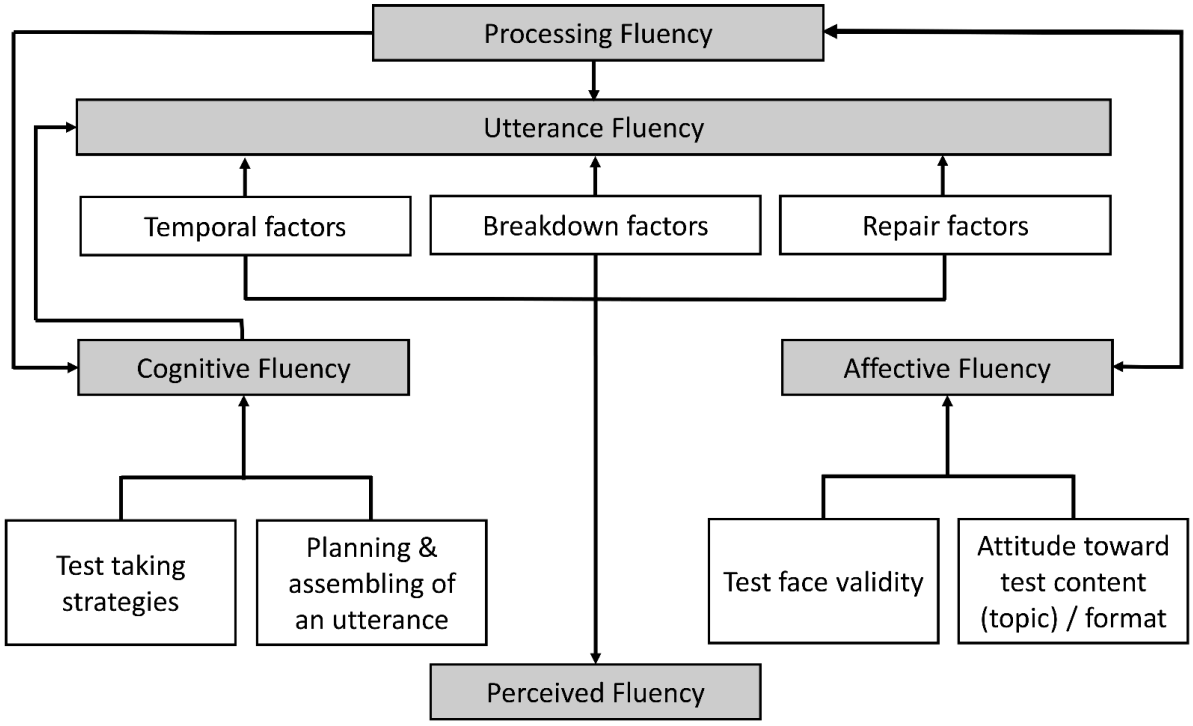

Figure 1 *Fluencies and measures theoretical relationships*

Spoken utterance fluency in direct and semi-direct oral proficiency interviews

Previous research investigating spoken utterance fluency has often been underpinned by examination of discourse strategies which are the procedures adopted by test takers to cope with the communicative demands of the testing event (Shohamy, 1994; O'Loughlin, 2001). For these early studies, discourse strategies primarily relate to the breakdown and repair measures of spoken utterance fluency. This is due to less accurate methods of calculating temporal measures being available until relatively recent times. However, test taker discourse strategy use is likely to affect all measures of their utterance fluency performance.

During a posteriori analysis of test taker output from a Hebrew direct and semi-direct speaking test, Shohamy (1994) reported that test takers used a greater variety of discourse strategies during the OPI, including hesitation and self-correction. Shohamy also asserted that output from the semi-direct OPI was more likely to be disfluent by test takers' use of a larger number of self-corrections and hesitation and silence. This latter finding has been partially corroborated in Bijani's (2019) study comparing English language learners' performance on five different speaking assessment tasks (four of which were monologic) in both OPI and SOPI settings. Bijani found that, among discourse strategies, self-correction was significantly more frequent in SOPIs, whereas hesitation had similar occurrences across the two assessment formats.

Results from Zhou's (2008) study, comparing direct and semi-direct versions of two monologic tasks from the speaking section of the Global Test of English Communication (GTEC), somewhat differed from Shohamy (1994), by showing that test takers used more general repetition during the direct OPI. More recently, when comparing test taker output in SOPI and OPI versions of monologic tasks from the Test of Spoken English (TSE) preparation materials, Choi's (2014) results generally agreed with findings from O'Loughlin's (2001) comparative study, which used closely-matched monologic tasks from direct and semi-direct versions of the Australian Assessment of Communicative English Skills (*access*) test, by showing test takers used self-correction, and repetition or paraphrasing with

broadly similar frequency.

Shohamy, Shumueli, and Gordon (1991) found instances of hesitation and silence in equal number in the OPI and SOPI, yet Zhou's (2008) study showed that the elicited output from the SOPI contained more filled pauses. In contrast, Choi (2014), again, generally concurred with O'Loughlin (2001, p. 93) as "the discourse strategies employed[…] appeared to be very similar in the two versions". Nonetheless, Choi (2014) did note a salient difference in that there was a lack of long silences in the OPI. Other evidence of differences between direct and simulated modes of OPIs were found by Chang, Lee, & Lee (2018) who noted increases in speaker anxiety among participants who underwent the OPI with a human interlocutor compared to those who took the computer-based version. This illustrates that test mode may impact affective, as well as cognitive, factors.

As stated by Galaczi (2010), views on simulated and face-to-face speaking assessments should not understand the two modes to be competing perspectives. Advantages and disadvantages afforded by one mode over the other should be taken into account by examiners within the perspective of their individual contexts. Therefore, the question is not how to eliminate unwanted variables in mode of delivery or how to simulate true-to-life elicitation conditions, but what to expect in the spoken output from one mode or the other. Investigating processing, cognitive and affective fluency's effect on spoken utterance fluency in SOPIs and OPIs will help to inform that expectation.

Method

Research questions and study design

> RQ1: Does test takers' level of spoken utterance fluency markedly differ according to output retrieved in performances from identical simulated and face-to-face oral proficiency interviews?
>
> RQ2: How do processing, cognitive, and affective fluency influence test-takers' spoken utterance fluency performance in the respective simulated and face-to-face oral proficiency interviews?

To address these research questions, a counter-balanced study design (see Figure 2) was conceptualized. Participants were not informed of task nature or content before either of the sessions. No practice or recency effects were noted during the second administration of the test, and thus a within-subjects data analysis was completed. The standard computer-based Aptis General SOPI test was administered, and only minor necessary modifications to the original test's computerized interlocutor input were made for direct OPI delivery. The online practice version of the Aptis General speaking test was chosen as the base test for this study to enable immediate access to data for the post-test stimulated recall verbal report interviews (SRVR).

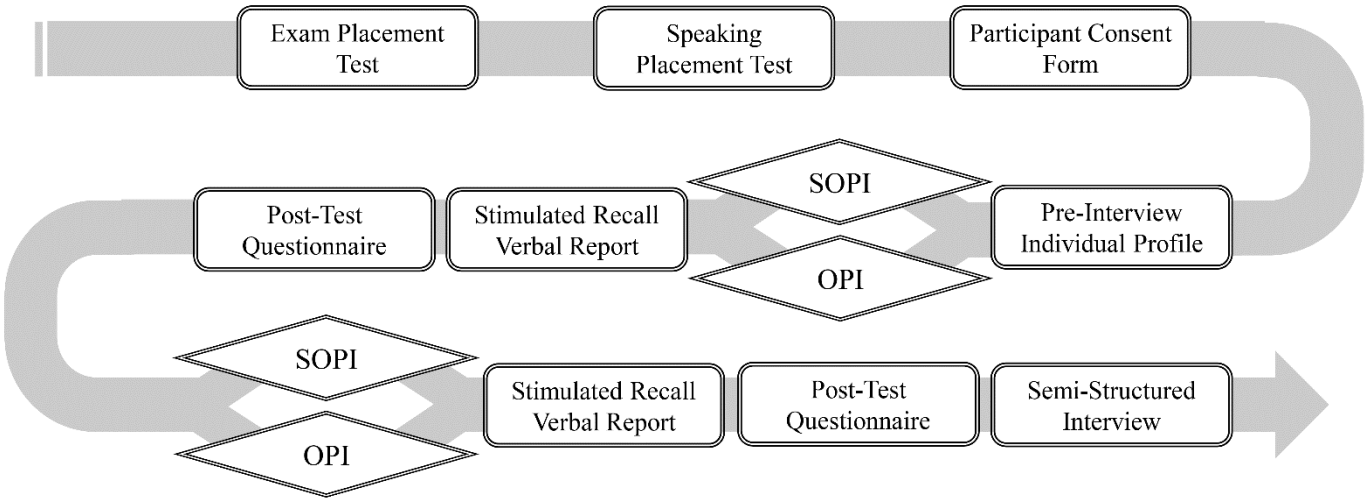

Figure 2 *Counter-balanced study design*

Aptis General speaking test

The Aptis General speaking test is a widely used speaking assessment delivered as a computer-based SOPI and consists of four parts using monologic tasks. Because of the monologic nature of the tasks, no managing interaction functions are reported as being measured. The test elicits increasingly more advanced language structures over timed sections which altogether take about 12 minutes to complete (see Figure 3).

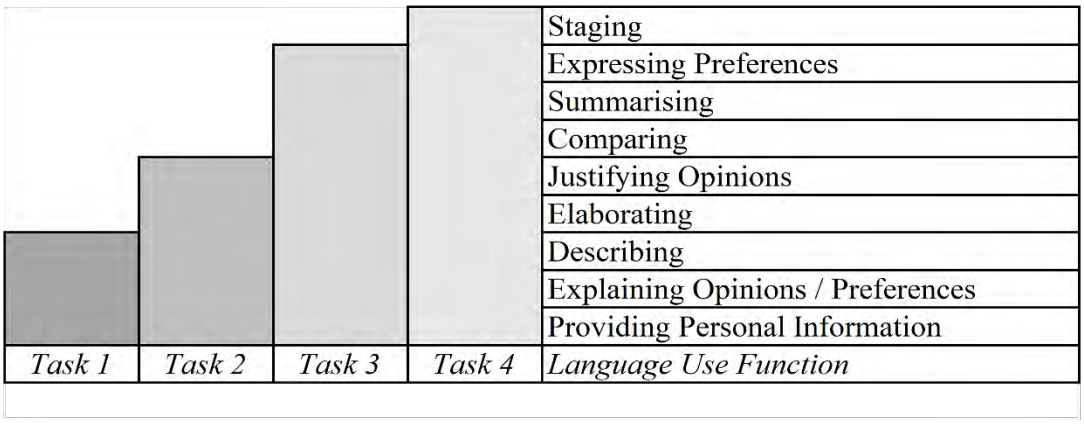

Figure 3 *Informational language functions (adapted from O'Sullivan* et al.*, 2020)*

Besides task equivalence in the general content knowledge and neutral cultural specificity across all four tasks, the features of the task input differ with regard to domain origin and the nature and complexity of the lexis used. O'Sullivan *et al.* (2020) state that the test's lexical levels are determined by word lists derived from the British National Corpus by Paul Nation (2006). The lists comprise 20 levels, each with 1,000-word families, with K1 referencing the most frequent 1,000-word families, and K2 the next most frequent 1,000-word families, and so on (see Table 1).

Table 1 *Input lexical feature differences (adapted from O'Sullivan* et al.*, 2020)*

Input lexical features	Task 1	Task 2	Task 3	Task 4
Lexical level	K1	K1, K2	K1, K2, K3	K1, K2, K3, K4
Nature of information	Concrete	Mostly concrete	Mostly concrete	Fairly abstract

Participants

Four participants were selected who completed the Cambridge English Language Assessment's (2020) Adult Learners' General English Exam Placement Test (EPT) that placed them as minimally intermediate English users. Face-to-face speaking placement tests were then conducted to ensure the EPT score validity, and all four participants were confirmed to be intermediate or higher, level L2 speakers of English. The four participants were assigned pseudonyms to prevent their identification as contributors to this study. Elena, Chloe, Emily, and Katrina were all females of Chinese ethnicity with a corresponding L1, who were aged between 21 to 35 years and self-reported as being highly familiar with computing and technology. The participants confirmed not having previously completed the online practice version of the test, and possessing no knowledge of the test's content, prior to the commencement of this study.

Procedure

The participants were randomly paired to create two groups which were assigned alternating test delivery modes for the two data collection sessions (see Table 2). A maximum of five days was allowed between the OPI and SOPI data collection sessions.

Table 2 *Delivery mode by participant and data collection session*

Group	Participant	Session 1	Session 2
1	Elena Chloe	Oral proficiency interview	Computer-based oral proficiency interview
2	Katrina Emily	Computer-based oral proficiency interview	Oral proficiency interview

The computer-based practice version of the Aptis General speaking test was administered in standard format in a computer laboratory and participants' spoken performances were captured by a DRD recorder placed on the computer desk. During task one, 30 seconds are provided to answer each question, and in tasks two and three, 45 seconds are provided per question. In task four, test takers have one minute to prepare to speak on a given prompt and then must speak for a total of two minutes.

The face-to-face OPI was identical to the computer-based version. Test delivery was conducted in a quiet room adjacent to the computer laboratory to minimize distractions. The SOPI visual test task prompts were made available for participants as appropriate during the OPI. The first author acted as examiner during the OPI and test performances were again captured by use of a DRD recorder. Administration of the speaking prompts in the face-to-face test mirrored the format and timing of the computer-based version.

The stimulated recall verbal reports (SRVR) were conducted in the same location as the OPIs and were completed directly after the administration of the speaking test in order to increase the likelihood of more accurate data being recalled. The protocols for conducting the SRVRs was adapted from Mackay and Gass (2000).

The second data collection session included a co-constructed, semi-structured interview conducted at the end of the session, allowing participants to expand on and contrast their responses in post-test questionnaires. Interviews were conducted in English and employed a recording device for later analysis. Furthermore, the exam content from the OPIs was available on a separate device during the interviews so that both participants and researcher were able to pause and reference the recorded exam content.

Analyses

The spoken output data from the OPIs of the four participants were transcribed following conversation analysis transcription notation conventions originally established by Atkinson and Heritage (1984) and later adapted by Lazaraton (2002).. To achieve accurate measurement of utterance fluency speed measures, Praat (Boersma & Weenink, 2013) software was used to simultaneously double-code the data. The following speed, composite, breakdown and repair output analyses for spoken temporal fluency were conducted:

(1) Speed measure
- (a) Articulation rate: total number of syllables divided by total amount of phonation time (excluding pauses) multiplied by 60

(2) Composite measures
- (a) Speech rate (pruned): total number of syllables divided by total performance time (including pauses) multiplied by 60.
- (b) Mean length of run (pruned): the mean number of syllables between two pauses

(3) Breakdown measures
- (a) Frequency of filled pauses (per 60 seconds)
- (b) Frequency of all silent pauses (per 60 seconds)

(4) Repair measures
- (a) Frequency of false starts and reformulations (per 60 seconds)
- (b) Frequency of partial or complete repetitions (per 60 seconds)
- (c) Frequency of self-corrections (per 60 seconds)

Data collected from the SRVRs and semi-structured interviews were not transcribed. These data were analyzed manually. As spoken output from the tests was available during the interviews, participants and the researcher were able to directly reference specific datum when asking or answering questions.

Results and Discussion

Participants performance was firstly measured for speed by analyzing the articulation rate of output (see Table 3).

Table 3 *Participant articulation rate per task and test performance*

Participant	Task 1 CB-OPI	OPI	Task 2 CB-OPI	OPI	Task 3 CB-OPI	OPI	Task 4 CB-OPI	OPI	Mean CB-OPI	OPI
Elena	225.48	203.47	182.47	188.44	206.74	186.04	187.56	177.16	203.16	191.1
Chloe	157.76	168.23	175.24	156.16	137.48	149.83	119.27	124.46	153.07	154.71
Katrina	127.14	129.8	118.94	127.67	105.5	131.17	93.73	121.49	114.85	128.74
Emily	158.03	184.58	160.49	165.09	166.08	118.48	158.76	142.38	161.26	164.68
Mean	167.1	171.52	159.28	159.34	153.95	146.38	139.83	141.37	158.08	159.8

The majority of participants' output across tasks and test overall exhibited a minimally higher articulation rate in the OPI version of the test. However, it is interesting to note that Elena's articulation rate was higher throughout the majority of the SOPI test task delivery than in the OPI equivalent. Affective fluency was likely an influence on this, because during the subsequent stimulated recall verbal report (SRVR), Elena ventured this was due to "something pushing her to speak" during the SOPI test. Chloe and Katrina provided further clarification to this by stating that they were very aware of the permitted response time during the SOPI because a countdown timer was present on the screen, and this in some form shaped their response to the task prompts and the speed at which it was delivered, which can be viewed as test-taking strategy and related to cognitive fluency.

Katrina stated that she felt that access to a countdown timer helped her to perform better during the SOPI delivery of the test and therefore this delivery mode had higher face validity for her. Furthermore, Elena mentioned that she attempted to interpret audial and visual clues as to the remaining time for each response which included interlocutor silence and glancing at the stopwatch during the OPI, yet she was wary that time and effort spent doing this may have affected her OPI performance. Worthy of attention was also the fact that participants' articulation rate mean values for both delivery modes decrease from task one to three which may demonstrate the effect of additional cognitive load as they progress through the tasks from the more concrete to abstract topics, with higher lexical levels and a wider range of language functions required from output.

Two measures of composite fluency were then analyzed. Firstly, participants' speech rate was measured with the resulting means by task and test provided in Table 4. Speech rate was affected similarly to articulation rate in that it often slowed as participants progressed through test tasks one to four. The mean speech rate for performances during the OPI version of the test clearly illustrate this; whilst the equivalent SOPI conforms to this pattern, albeit the speech rate recorded in test task three being seen as an individual outlier. Overall, the OPI delivery was evidenced as eliciting test taker output with a higher speech rate, although the observed difference was minimal.

Table 4 *Speech rate task and test performance means*

	Task 1	Task 2	Task 3	Task 4	Test mean
SOPI	155.08	140.43	148.72	136.08	145.08
OPI	167.64	156.69	149.37	137.97	152.92

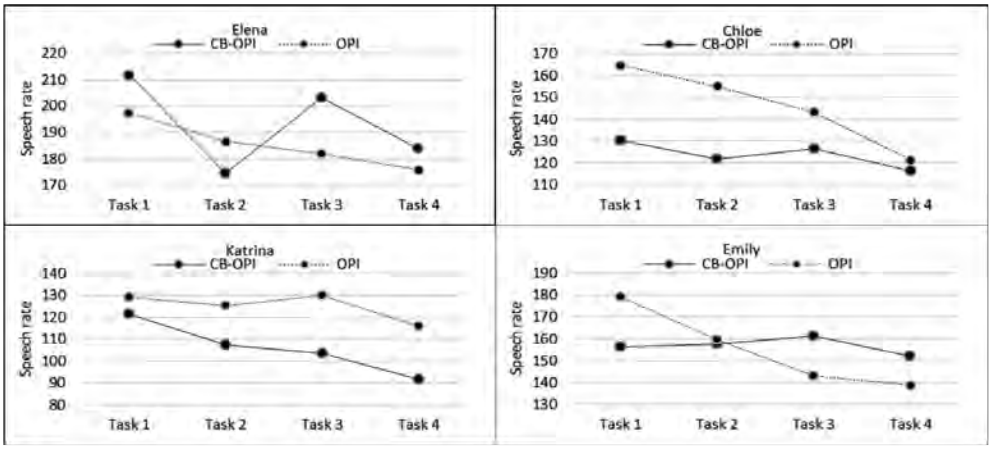

Figure 4 *Speech rate per participant and task*

Individualized speech rate graphs per participant are presented in Figure 4. The composite measure of speech rate appeared to be influenced primarily by test takers' cognitive fluency, because on listening to her performances for both delivery modes, Katrina described using the test taking strategy of lengthening syllables to avoid pausing, and the effect of this was evident in the post-test speech rate analysis. A more successful cognitive test taking strategy in terms of all measures of spoken utterance fluency was Elena's use of the lexical filler *like* to accomplish a similar outcome to that of Katrina, although devices such as this can be frowned upon by marking rubrics as colloquial and not being indicative of educated English language speaking.

The second measure of composite fluency analyzed was mean length of run (see Table 5). Longer mean length of runs were evenly distributed between delivery modes with the first two test tasks favoring the OPI and the latter two the SOPI. Similar to speech rate, and the speed measure of articulation rate, the means for this second composite measure of fluency indicated marginally better fluency performance from participants in the OPI version of the test.

Table 5 *Mean length of runs per task and test performance*

	Task 1	Task 2	Task 3	Task 4	Test mean
SOPI	6.94	6.42	8.2	6.72	7.07
OPI	10.03	8.77	7.77	6.52	8.27

Graphs illustrating the mean length of runs per participant and task are presented in Figure 5. The longer mean length of runs produced in the OPI performances, in comparison to spoken output from the semi-direct tests, belied Elena and Emily's SRVR and semi-structured interviews' assertions. Harkening affective fluency, they maintained that a traditional face-to-face speaking test is a conversation between two people if it is accepted that a feature of conversation is its more informal register compared to institutional talk, with the latter containing more complex language and perhaps longer runs of speech. However, from the perspective of face validity, another indicator of affective fluency, Emily was certain that actually talking to someone in a face-to-face OPI meant that a rater's marking of her output would likely be more truly indicative of her spoken language proficiency, and her preference for an OPI is shared among a relatively large number of potential test takers (Qian, 2009). Remarkably, however, Emily felt more confident in retaking a SOPI than another OPI in the

future and maintained that a SOPI is still an appropriate test delivery mode that elicited output representative of her true speaking ability.

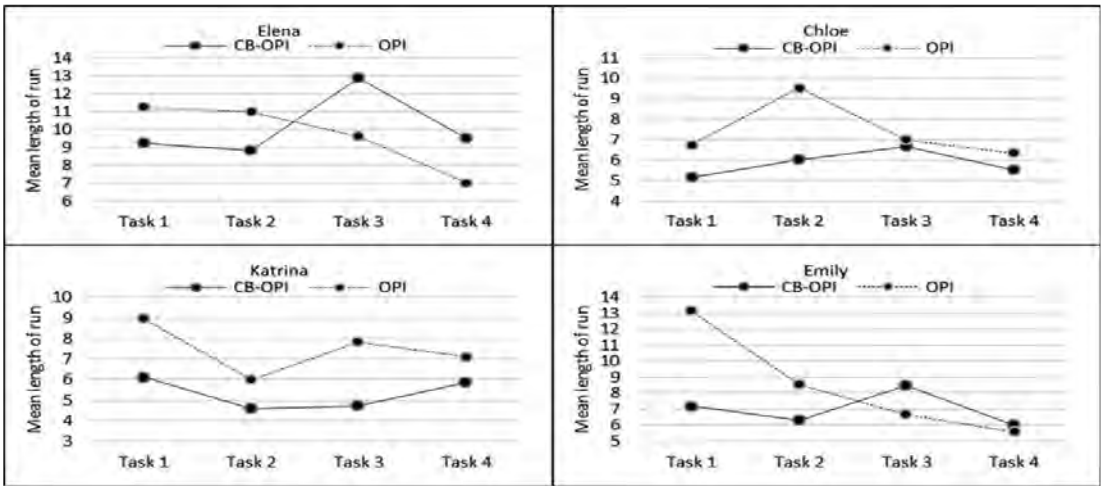

Figure 5 *Mean length of runs per participant and task*

The breakdown fluency measures of filled and unfilled pauses were calculated per minute of participant spoken performance time (see Table 7). Participants used a greater number of filled pauses throughout all tasks administered in OPI form, and saliently output from the SOPI exhibited that this measure of fluency was increasingly affected as participants progressed throughout tasks in this delivery mode.

Table 7 *Filled pauses per minute of spoken performance*

	Task 1	Task 2	Task 3	Task 4	Test mean
SOPI	8.02	8.4	8.69	8.77	8.47
OPI	10.37	10.66	10.35	9.82	10.3

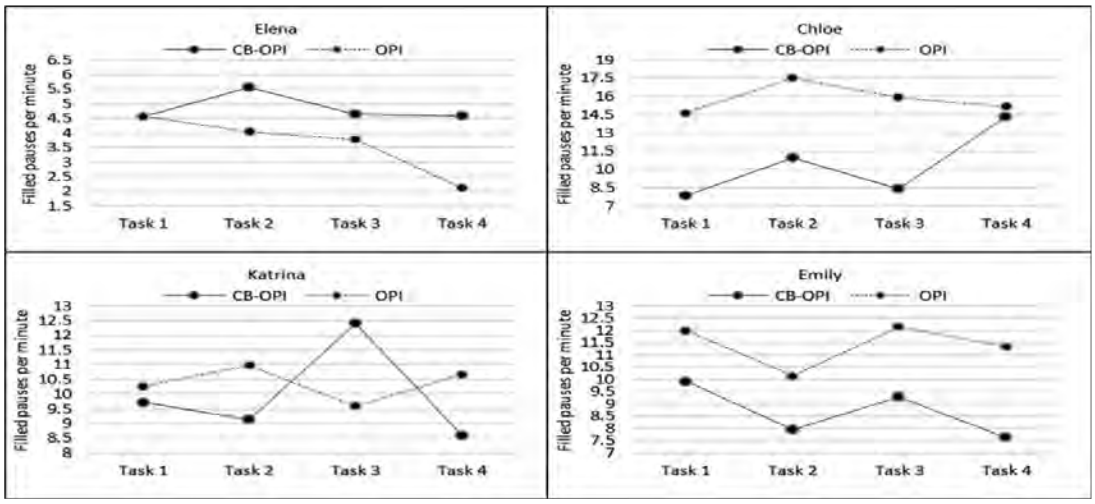

Figure 6 *Filled pauses per participant and task*

It is less certain how individual performances between delivery modes were affected in terms of filled pause use (see Figure 6), yet both cognitive and affective fluencies were noted as the underpinning reason for their use. In terms of cognitive fluency, all participants indicated in their respective SRVRs that filled pause use represented time spent searching for vocabulary and/or grammar, in contrast to thinking about task prompt topics. Affective fluency was also a factor as test takers' reported a perceived greater pressure to focus on accuracy in the OPI performance of the test. For Chloe, Katrina and Emily, the notion of OPI discourse as conversation was mentioned in their SRVR sessions and this was posited by them as the most likely explanation for their increased filled pause use in this test delivery mode. Unfortunately, further clarification of OPI filled pause use was not forthcoming during their SRVRs, although proposals such as to maintain turn (response), or avoid uncomfortable silence are somewhat likely explanations.

The second breakdown fluency measure to be analyzed was unfilled pause use. Results were atypical of those for filled pause use. Unfilled pauses were used more frequently in the SOPI across all tasks and the test overall (see Table 8).

Table 8 *Unfilled pauses per minute of spoken performance*

	Task 1	Task 2	Task 3	Task 4	Test mean
SOPI	23.37	24.32	19.79	23.47	22.74
OPI	18.68	17.43	19.32	21.02	19.11

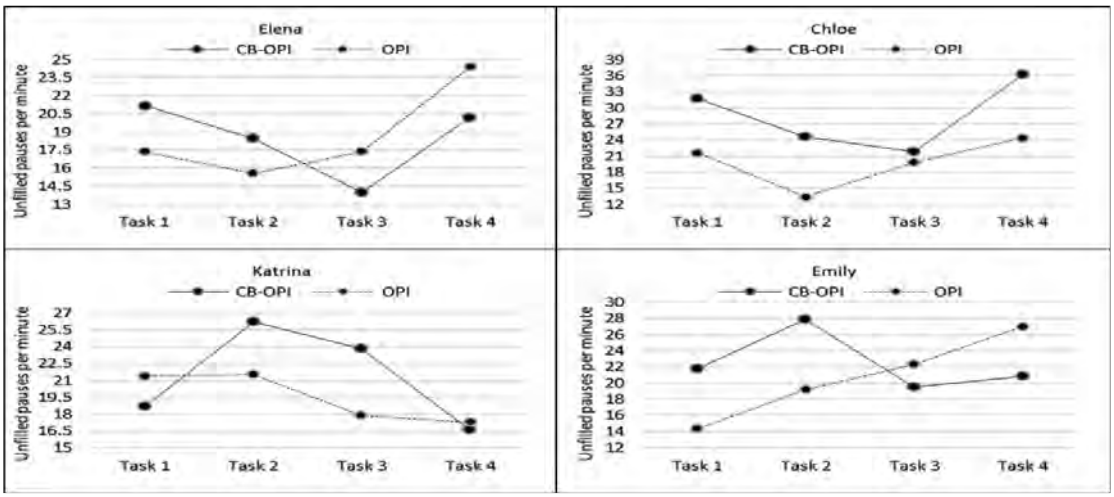

Figure 7 Unfilled pauses per participant and task

Figure 7 presents graphs of unfilled pause use by participant and task. Distinguishing patterns of use for individual participants were for the most part absent. Evidence of breakdown through unfilled pauses according to the staged increasing difficulty of tasks was only found in Emily's OPI performance (see Figure 6). Affective fluency was an influencer of unfilled pause use, as Katrina stated that due to her being unable to read any non-verbal clues from the SOPI interlocutor, more nervousness was experienced because both conscious and subconscious feedback from an interlocutor helps a test taker assess the success of conveyed meaning and adjust their response to the clues communicated. The use of filled and unfilled pauses was the spoken fluency measure that was most likely subject to differ between delivery modes. Furthermore, the results from this present study were similar with those of Zhou (2008), by showing that the elicited output from the semi-direct test

contained more unfilled pauses.

Three fluency repair measures were analyzed in output from the delivery modes including repetitions, self-corrections, and reformulations (see Table 9). Repetition was deemed as such if it referenced test takers' previous output and not interlocutor input. Only if an entire identical lexical item or phrase was consecutively repeated was it classified as repetition, which avoided possible ambiguities with test takers' self-correction. Repetitions for stylistic effect were also discounted. Each occurrence of repetition was calculated and not the number of items repeated. The number of reformulations in expressing ideas and self-corrections were counted. Reformulations were differentiated from self-correction by the participants' interruption of speech to add a new proposition, instead of doing so to correct an error, and those that expressed lexical items (stor-shops) were not included.

Table 9 *Repetitions, self-corrections and reformulations per minute of test spoken output*

	Total repairs		Repetitions		Self-corrections		Reformulations	
Participant	**SOPI**	**OPI**	**SOPI**	**OPI**	**SOPI**	**OPI**	**SOPI**	**OPI**
Elena	2.511	2.22	0.279	0.555	2.093	1.388	0.139	0.277
Chloe	1.584	2.142	0.792	1.377	0.528	0.612	0.264	0.153
Katrina	5.301	5.052	2.533	2.721	2.109	1.684	0.659	0.647
Emily	4.494	4.988	3.211	3.464	1.027	1.247	0.256	0.277

More word and phrasal level repetition was found in all performances of the OPI version of the test. This finding replicated that of Zhou (2008) who found that repetition was used more in the direct test, irrespective of test taker proficiency. Yet, this finding was inconsistent with O'Loughlin (2001) and Choi (2014)'s results. A likely explanation for this study's repetition results' being similar to those of Zhou (2008) is that monologic tasks and a similar adherence to strictly controlled interlocutor input were used in both studies. The latter refers to an increased degree of interactiveness in test taker–interlocutor OPI discourse that could well be significant for the repair measures of utterance fluency, because test takers may "experience a higher level of nervousness when facing an interlocutor, which makes them hesitant and consequently causes them to use more repetition" (Zhou, 2008, p.202). Yet, this was not the primary reason for repetition use by Chloe and Emily because data retrieved from the SRVRs suggested that repetition was a device they equally used to bide thinking time for the searching of appropriate vocabulary and grammar for upcoming output.

The results of the self-correction and reformulation analyses were devoid of patterning at the level of test. However, Katrina's performance in terms of self-correction conformed to her suggestion in the semi-structured interview that affective fluency determined use as there was a need for additional self-correction during an OPI because it was necessary to make a live interlocutor understand her awareness and ability to correct errors. Shohamy's (1994) and Bijani's (2019) claim that self-correction is more likely to be found in the semi-direct mode appears to be problematic, as the results of this study along with Koike (1998), O'Loughlin (2001), and Choi (2014) have found otherwise. Nonetheless, universal agreement is found between this study and all other previous related research, in that it is clear the number of reformulations was not influenced by mode effect.

Conclusion

The purpose of this study was to investigate qualitative explanations for test taker utterance fluency in spoken output from an identical direct and semi-direct speaking test. In reference to RQ1, it is evident that test taker spoken utterance fluency may well be affected by the computer-based delivery of speaking tests, because the main difference observed was influenced by mode effect. This difference in spoken utterance fluency primarily manifested itself through test takers' use of filled and unfilled pauses. With this premise, a conclusion that can be drawn is that direct delivery can cause more vocal hesitation, and semi-direct delivery more non-verbal hesitation, for test takers. For interpretation of this finding, it is vital that the minimal degree of this differentiation should be remembered, because although it was certainly observable in output analyses, it was highly unlikely that these differences would impact a test takers' score on taking the test used in this study, as they would almost certainly be unobservable to an examiner uponlistening to the performances and also a live interlocutor present for a direct OPI delivery.

In response to RQ2, the qualitative data retrieved from this study implies that affective fluency plays a role in test takers' processing fluency that in turn impacts spoken utterance fluency performance. From the aspect of cognitive fluency, the test takers proposed that the searching for appropriate vocabulary and grammar for upcoming language output influenced all spoken utterance fluency measures, and therefore indirectly affected perceived fluency. Furthermore, the test-taking strategy of lengthening syllables to provide continuity in spoken utterance fluency was also used and reflects the cognitive aspect of processing fluency. In regard to affective fluency, test takers' suggested that the (non)presentation of available response time affected their processing fluency and speed of delivery measure of utterance fluency. Other notable markers of affective fluency influencing processing fluency and the test takers' resulting spoken utterance fluency were attitudes toward a test delivery mode's face validity, along with the test takers' feelings toward a live interlocutor being present during the OPIs.

Limitations and future research directions

Several limitations should be remembered when interpreting the results of this study. Firstly, this study was necessarily sample dependent, which prevents the generalizability of findings, because the participants were from a narrowly defined population, which is unlikely to be fully representative of a broader L2 English user population. Furthermore, the relatively strict control of interlocutor input in the administered OPIs in this study means attempts at generalizing findings to output from other task types or test events with a higher degree of interactiveness is not possible. Moreover, this study used identical task type and content, which may have encouraged recency effect in the participants' performances during the second administration of the test. And finally, it is also feasible that if a female had acted as the interlocutor in the OPIs, test takers spoken utterance fluency performances might have differed (Porter & Shen, 1991; O'Sullivan, 2000).

Future related mixed method studies, with the qualitative data building upon the quantitative, would be beneficial for better understanding cognitive and affective fluency's role in determining spoken utterance fluency through the possible mediation of processing fluency. Larger-scale research that examines test taker characteristics and test taking strategies, in addition to the interactional context of delivery mode, may deliver further insight as to how these affect spoken utterance fluency performance in direct and semi-direct speaking tests, through test takers' processing, cognitive, and affective fluencies. Results from this research could enable test developers, material writers, and exam preparation teachers the opportunity to provide positive washback in classrooms, so as to help ensure test takers are adequately prepared for their direct or semi-direct speaking test performance,

including, for example, the introduction of increased practice of recorded spoken monologic tasks, which can be listened to and self-reflected upon by potential test takers and commented on by exam preparation teachers.

References

Alderson, C. (2004). The shape of things to come: will it be the normal distribution. In M. Milanovic, & C. Weir, (Eds.), *Studies in language testing 18: European language testing in a global context* (pp. 1–26), Cambridge: Cambridge University Press.

Atkinson, J.M., & Heritage, J. (1984). *Structures of social action: Studies in conversational analysis*. Cambridge: Cambridge University Press.

Bijani, H. (2019). Evaluating the effectiveness of the training program on direct and semi-direct oral proficiency assessment: A case of multifaceted Rasch analysis. *Cogent Education,* 6(1), 1–20. https://doi.org/10.1080/2331186X.2019.1670592

Boersma, P., & Weenink, D. (2013). *Praat: doing phonetics by computer* [Computer program] Version 6.1.16. Retrieved from http://www.praat.org/

British Council. (2020). *Aptis General Speaking Test: Practice version*. Retrieved from https://www.britishcouncil.org/aptis-practice-tests/AptisSpeakingPractice/

Cambridge English Language Assessment. (2020). *Test your English–Adult Learners*. Retrieved from http://www.cambridgeenglish.org/test-your-english/adult-learners/

Chang, S. L., Lee, S. -D., & Lee, S. -P. (2018). Divergent effects of direct and semi-direct oral assessment on psychological anxiety and physiological response in EFL college students. *Asian EFL Journal Research Articles, 20*(12), 252–269.

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

Choi, I. (2014). The comparability of direct and semi-direct oral proficiency interviews in a foreign language context: a case study with advanced Korean learners of English. *Language Research, 50*(2), 545–568. https://doi.org/10371/93290/1/12

De Jong, N.H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3), 237–254. https://doi.org/10.1080/15434303.2018.1477780

De Jong, N.H., Steinel, M.P., Florijn, A., Schoonen, R., & Hulstijn, J. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*(5), 893–916. https://doi.org/10.1017/S0142716412000069

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: challenges and opportunities. In Araújo, L. (Ed.) *Computer-based assessment (CBA) of foreign language speaking skills* (pp. 29–51), Luxembourg, LU: Publications Office of the European Union. https://doi.org/10.2788/30519

Graf, L.K., Mayer, S., & Landwehr, J.R. (2018). Measuring processing fluency: one versus five items. *Journal of Consumer Psychology, 28*(3), 393–411. https://doi.org/10.1002/jcpy.1021

Huensch, A., & Tracy-Ventura, N. (2016). Understanding second language fluency behaviour: the effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics, 38*(4), 755–785.

Jaud, D. A., & Melnyk, V. (2020). The effect of text-only versus text-and-image wine labels on liking, taste and purchase intentions. The mediating role of affective fluency. *Journal of Retailing and Consumer Services, 53*, 1–10.

Koike, D. A. (1998). What happens when there's no one to talk to? Spanish foreign language discourse in simulated oral proficiency interviews. In R. Young, & A. Weiyun He (Eds.),

Talking and testing: Discourse approaches to the assessment of oral proficiency. Philadelphia, PA: John Benjamins. https://doi/org/10.1075/sibil.14.06koi

Lazaraton, A. (2002). A qualitative approach to the validation of oral language tests. *Studies in Language Testing, 14*. Cambridge: Cambridge University Press.

Luoma, S. (1997). *Comparability of a tape-mediated and face-to-face test of speaking*: *a triangulation study*, Unpublished licentiate thesis, Finland: University of Jyvaskyla. Retrieved from http://urn.fi/URN:NBN:fi:jyu-1997698892

Mackay, A., & Gass, S.M. (2000). *Stimulated recall methodology in second language* r*esearch*. Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal, 1*(1), 1–15. https://doi.org/10.1017/S2041536210000012

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82.

O'Loughlin, K. (2001). The equivalence of direct and semi-direct speaking tests. *Studies in language testing, 13*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *Journal of System, 28*, 373–386. https://doi.org/10.1016/S0346-251X(00)00018-X

O'Sullivan, B. (2015). *Aptis test development approach, Aptis Technical Report* (TR/2015/001). Retrieved from https://www.britishcouncil.org/exam/aptis/research/publications

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K, (2020). *Technical Report: Aptis General Technical Manual Version 2.2* (TR/2020/001). Retrieved from https://www.britishcouncil.org/exam/aptis/research/publications

O'Sullivan, B., & Weir, C.J. (2011). Language testing and validation. In B. O'Sullivan, (Ed.), *Language testing: theories & practices* (pp.13–32), Basingstoke: Palgrave Macmillan.

Peltonen, P. (2017). Temporal fluency and problem-solving in interaction: an exploratory study of fluency resources in L2 dialogue. *System, 70*, 1–13. https://doi.org/10.1016/j.system.2017.08.009

Porter, D., & Shen. H. (1991). Sex, status, and style in the interview. *Dolphin, 21*, 117–128.

Qian, 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. (2018). Output register parallelism in an identical direct and semi-direct speaking test: A case study. *International Journal of Computer-Assisted Language Learning and Teaching, 8*(2), 75–91. https://doi.org/10.4018/IJCALLT.2018040105

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*(4), 364–382.

Sato, M. (2014). Exploring the construct of interactional oral fluency: second language acquisition and language testing approaches. *System, 45*, 79–91. https://doi.org/10.1016/j.system.2014.05.004

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching, 54*, 79–95.

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., Shumueli, D., & Gordon, C. (1991). *The validity of concurrent validity of a direct vs. semi-direct test of oral proficiency*, a paper presented at the 13th Language Testing Research Colloquium. United States of America: LTRC.

Weir, C. J. (2005). *Language testing and validation: an evidenced-based approach*. New York,: Palgrave Macmillan.

Zhou, Y. J. (2008). A comparison of speech samples of monologic tasks in speaking tests between computer-delivered and face-to-face modes. *Japan Language Testing Association Journal, 11*, 189–208. https://doi.org/10.20622/jltaj.11.0_189

Zhou, Y. J. (2015). Computer-delivered or face-to-face: effects of delivery mode on the testing of second language speaking. *Language Testing in Asia, 5*(2), 1–16. https://doi.org/10.1186/s40468-014-0012-y