Chapter 10
Toward the Future of Computer-Assisted Language Testing:
Assessing Spoken Performance Through Semi-Direct Tests

Ethan Douglas Quaid
https://orcid.org/0000-0003-2682-2347
Xi'an Jiaotong-Liverpool University, China

Alex Barrett
https://orcid.org/0000-0003-1229-9743
Xi'an Jiaotong-Liverpool University, China

ABSTRACT

Semi-direct speaking tests have become an increasingly favored method of assessing spoken performance in recent years. Underpinning evidence for their continued development and use has been largely contingent on language testing and assessment researchers' claim of their interchangeability with more traditional, direct face-to-face oral proficiency interviews through theoretical and empirical investigations from multiple perspectives. This chapter initially provides background and research synopses of four significant test facets that have formed the bases for semi-direct and direct speaking test comparison studies. These are followed by the inclusion of a recent case study comparing test taker output from a computer-based Aptis speaking test and a purposively developed identical face-to-face oral proficiency interview that found a slight register shift which may be viewed as advantageous for semi-direct speaking tests. Finally, future research directions are proposed in light of the recent developments in the semi-direct speaking testing research presented throughout this chapter.

DOI: 10.4018/978-1-7998-1282-1.ch010

INTRODUCTION

Computer-based (CB) and computer-mediated (CM) testing is today, more than ever, being seen as the future of language assessment, primarily through the former's increasing role in testing contexts. The broadening use of computer-based oral proficiency interviews (CB-OPI) by test developers has necessitated that myriad studies be conducted by language testing and assessment researchers in order that this delivery mode is underpinned by sound empirical knowledge. This research has provided a growing body of evidence suggesting a degree of interchangeability between CB and face-to-face oral proficiency interviews (OPI). This chapter firstly reviews CB speaking tests' comparability with OPIs by providing synopses of background and current research findings pertaining to four of the most significant test facets supporting arguments for the use of semi-direct speaking tests over their direct counterparts.

Practicality is the first of these test facets and arguably the most important from a test developers and administering institution's perspective. Representing the second test factor is face validity which has proven to be instrumental in the decision-making process when considering the development and use of semi-direct speaking tests. Test reliability and concurrent validity are essential test measurement factors, and hence have been chosen as the latter two facets to be included within this chapter examining direct and semi-direct speaking test interchangeability. Research studies encapsulating these four test facets have often promoted the development and use of semi-direct speaking tests, and this has led to concerns that the important factor of test taker spoken output from performances has often been overlooked (Shohamy, 1994; Weir, 2005; Zhou, 2015).

The second half of this chapter examines and contrasts test taker spoken output in direct and semi-direct speaking test delivery modes by the use of a recent case study investigating the equivalence of spoken language register in test event performances. This exemplified case study provides additional empirical evidence supporting the interchangeability of test taker output register in direct and semi-direct speaking tests, by comparing spoken output elicited from four Chinese first language users of intermediate level spoken English in a British Council *Aptis General* CB-OPI and a purposively designed identical face-to-face direct OPI version. The results from this differentiation provide additional empirical evidence contributing toward establishing the equivalence of CB and face-to-face test delivery modes in a foreign language speaking test context. In particular, the validation of CB input as an elicitation instrument through interpretation of performance.

DIRECT AND SEMI-DIRECT SPEAKING TESTS

Clark (1979) labels direct speaking tests as consisting of "procedures in which the examinee is asked to engage in a face-to-face communicative exchange with one or more human interlocutors" (p.36). However, a truly direct test measures proficiency inside the identified target language use (TLU) domain, and therefore it is preferable to identify test directness on a cline between the indirect structural model and live performance in the TLU domain as the most direct. Mirroring Clark's (1979) communicative exchange, Luoma (1997) states "the main characteristic of the live test mode is that interaction in it is two-directional", and asserts that "the construct assessed is clearly related to spoken interaction" (p. 44). Although these suppositions were likely to be a given, when early theoretical models underpinned language testing, today the relatively scripted nature of many speaking tests, especially those termed as being high stakes, may well have rendered them problematic. Thus, Fulcher's (2014) view that direct

testing only implies that *direct* involves live physical interaction with a human interlocutor and nothing more will be applied, and henceforth a direct test will be represented by a face-to-face OPI throughout the remainder of this chapter.

Semi-direct speaking tests are defined for the purpose of this chapter as tests that elicit spoken performance by means of non-human or technology-assisted elicitation procedures. This description encompasses both the early audio tape-based simulated oral proficiency interviews (SOPI) and today's CB mode of testing, as well as CM test input found in the mode of video-conferencing. Much **c**riticism has been leveled at semi-direct testing from language test developers, researchers and users alike, and thus merely establishing some form of parity between delivery modes has been a primary concern for the field of language assessment and testing. Nonetheless, CB and CM speaking tests' increased practicality is a test quality that most literature argues is supportive of these delivery modes.

PRACTICALITY

Practicality is the extent to which resources support the design and development of the test as well as its ongoing use. A semi-direct test certainly provides a more superior practical alternative, because of the "flexibility with which it can be administered" (Dooey, 2008, p. 30). Semi-direct speaking tests can be conducted with large groups of candidates completing the test simultaneously meaning a shorter period of time is generally needed for administration of the test (Yonezaki, 2016); whereas face-to-face OPIs often necessitate additional procedural tasks, that require supplemental human, material, and time resources. Semi-direct testing can also be "especially pertinent in foreign language testing situations in which it is difficult to establish a reliable pool of qualified interviewers" (Choi, 2014, p. 546).

In an increasingly automated world, issues in video and audio quality are less of a concern for developers and administrators of CM language tests, due to any breakdown of communication being the exception rather than the norm. For example, studies (Kim & Craig, 2012; Loranc-Paszylk, 2015) investigating videoconferencing as a semi-direct delivery mode for language testing have found that in terms of delivery this mode is stable enough to be considered an alternative to more traditional face-to-face direct input. Additionally, Nakatsuhara, Inoue, Berry, and Galaczi (2017a) have asserted that while reducing many of the practical challenges for speaking test delivery, the CM videoconferencing delivery mode does not subvert the construct of interaction of a face-to-face OPI. However, this is not a universally agreed position for all semi-direct modes of delivery as Brooks and Swain (2014) state that CB-OPIs can limit interaction, and this could then affect the predictive validity of performance during tests to performance in TLU domains.

The practicality of CB and CM testing also benefits test takers by allowing them to complete a spoken language proficiency test, needing only access to a computer connected to the internet. This provides a wider constituency of test takers whom may be located in remote geographical areas, where it is impossible to conduct OPIs, with accessibility to a test (Nakatsuhara et al., 2017a). This inclusiveness encourages equal opportunity and fairness to be realized for all potential test takers which are important factors not to be overlooked in the social dimension of language testing.

Practicality advantages of CB speaking tests have been reported by Chinese *Aptis* CB-OPI test takers with the test being found to be "easy and convenient to operate" (Zheng & Zhang, 2017). Moreover, an interesting recent study exploring the use of the CM video-conferencing delivery mode for the Interna-

tional English Language Testing System (IELTS) also reported participants finding the test easier than the long-standing traditional OPI equivalent (Berry, Nakatsuhara, Inoue, & Galaczi, 2018).

Institutional test score users are important stakeholders for whom CB and CM speaking tests can also provide distinct practicality advantages. Perhaps the most significant of these is that test results can be made available to them faster, enabling program entry decisions to be made in a timelier manner, which is increasingly important in the now internationalized education sector. Likewise, decision-making for migration and work purposes is able to be completed in a timelier manner, allowing government bodies and employers to streamline administrative processes.

Evidently, there are numerous practicality benefits for the continued implementation and use of semi-direct speaking tests; however, possible risks to their feasibility cannot be discounted. These include test takers' lack of computer familiarity leading to construct-irrelevant variance and possible test security issues. Furthermore, the procedural tasks, which require additional administration to complete, and the use of human, material and time resources in conducting face-to-face OPIs may well be insignificant when considering the alternative of developing or purchasing a CB speaking test, which could be especially prohibitive in terms of monetary cost for smaller institutional test score users.

These issues notwithstanding, a practicality advantage has led to ever-increasing support for the continued introduction of new semi-direct speaking tests, and encourages established tests of both high and low stakes denomination to replace their direct OPI with the CB or CM delivery mode. However, Weir's (2005) heretical view that "practicality is simply not a necessary condition for validity" (p. 49) leads us to examine three further test factors, which could demonstrate semi-direct tests' parity with their direct OPI counterparts, beginning with face validity.

FACE VALIDITY

Face validity refers to how test developers, institutional test score users and associated stakeholders, in addition to the test takers themselves, perceive the test to be an accurate and fair measure of spoken performance, as well as their affective variables toward a test delivery mode more generally. Interchanging a live interlocutor with a computer may seem counterintuitive for a proficiency assessment measuring speaking ability, and hence the face validity of CB-OPIs is often seen as being key to their success. In contrast to a recent study conducted by Alutumi (2018) that suggests positive attitudes were indicated towards integrating computer-assisted language learning for TOEFL iBT exam preparation training, research findings have shown mixed results as to how test takers feel about actually using computers during oral assessment procedures.

An early study by Kenyon and Malabanga (2001) compared test taker attitude toward both a SOPI and CB-OPI and compared them with their affective variables in direct OPIs. Participants who took both the SOPI and the CB-OPI found the CB exam to be less difficult due to its adaptive nature, but otherwise both tests scored similarly. The two semi-direct tests were then compared to an OPI and results indicated that participants felt the differently mediated exams to be similar. Despite this likeness, participants in this study felt the face-to-face OPI was more personal and consequently inspired language that was more conversational and thereupon more closely aligned with *real-life* speaking scenarios.

A later study by Surface, Poncheri, and Bhavsar (2008) included a comparative question of preference between direct and semi-direct oral proficiency tests and their results indicated a predilection for the OPI, although it was speculated that this was due primarily to respondents' dislike of an avatar used in

the CB-OPI and less to do with skepticism concerning the validity of the test. Test takers' predilection for OPIs was corroborated by Qian (2009) who measured test-taker preference between the IELTS OPI and an in-house semi-direct OPI at the Hong Kong Polytechnic University. Furthermore, a more recent study by Yonezaki (2016) demonstrated a definitive preference for the direct test. Test taker comments elicited centered on the preference for a speaking test that was similar to natural daily life interaction, although a few respondents mentioned a preference for the semi-direct test as they felt more nervous when confronted with a live interlocutor.

Zhan and Wan (2016) measured test taker opinions of a semi-direct OPI used in China which showed that participants felt "their pragmatic and strategic competencies in communication were not measured" (p. 373), because the CB test seemed inauthentic. It should be noted though that Zhan and Wan (2016) were not comparing their CB-OPI to an OPI, and thus participants did not have the benefit of an explicit contrast. Lastly, Dunlea et al. (2018) measured test-taker preference in Vietnam between face-to-face and CB tests in their examination of concurrent validity between the *Aptis* and a similar non-computerized OPI, finding a strong preference for the face-to-face test as well, with 74% of respondents agreeing or strongly agreeing to a preference for it.

The evidence seems to indicate a preference by test takers for OPIs due to reasons of authenticity and natural speech production which it is perceived can only arise from human-human interaction. This seems to be incongruent with seminal research in human-computer interaction which testifies that people tend to interact with computers similarly to how they would with people (Reeves & Nass, 1996). Despite this, examining test taker affective variables through general preference toward direct and semi-direct tests is just one aspect of face validity.

A second important characteristic of face validity is the determination of whether a test taker feels an assessment fairly and accurately measures the constructs it purports to measure. For example, Amengual-Pizarro and Garcia-Laborda (2017) discovered that test takers who took both direct and semi-direct OPIs found that the CB-OPI was a perfectly valid measure of oral competence. This perception of the CB-OPIs validity was due to the study participants noting that the CB-OPI was "convenient, intuitive, and user-friendly" (p. 32).

Zheng and Zhang (2017) measured test taker attitudes to both a CB test and a paper based test that measured the four major skills and revealed that test takers found advantages and disadvantages to both formats without a clear preference for one over the other, although it was discovered that the speaking portion of the CB test was the most difficult for participants. On the other end of the spectrum, a study conducted by Stricker and Attali (2010) which measured attitudes toward the *TOEFL iBT* found the speaking section of the exam garnered many more unfavorable opinions compared to the sections of the test assessing the other three language use skills.

Overall, test takers tend to feel that face-to-face OPIs are preferable. Therefore, continued development and adoption of semi-direct speaking tests may encounter some resistance with test takers who feel that a more accurate assessment of their speaking can only be achieved in direct OPIs. As artificial intelligence continues to develop, a CB-OPI which can adapt and perhaps personalize the experience of CB-OPIs, without compromising test reliability, could well be a valid solution.

RELIABILITY

Test reliability is the consistency of test performance measurement. It is typically examined in two ways: test-retest reliability (consistency in test performance when a test taker takes the same test twice) and rater agreement (consistency in intra- and inter-rating). Thompson, Cox, and Knapp's (2016) study found a high degree of reliability in both test-retest and rater agreement for an OPI and CB-OPI, which confirmed results from previous studies (Abadin, Fried, Good, & Surface, 2012; Surface, Harman, Watson, & Thompson, 2009; Surface, Poncheri, & Bhavsar, 2008).

More recent studies examining the scoring reliability of high profile CB-OPIs have underscored this delivery mode's consistency in the rating of test taker performance. For instance, Schmidgall's (2017) study analyzed the reliability of the Test Of English for International Communication (TOEIC) speaking scores across different levels of the scoring procedure using the G-theory framework and found the dependability of the TOEIC Speaking scale scores was relatively high.

Despite the potential test-retest reliability of both OPIs and CB-OPIs, detractors have put forth valid points, regarding construct irrevelant variance. Lazarton's (1996) study examining the types of linguistic and interactional support that an interlocutor provides to the test taker in an OPI found that the "potential for uneven interview performance in a face-to-face interview is one reason that [semi-direct tests] are so appealing i.e. they remove the variability that a live interlocutor introduces" (p.154).

An example of interlocutor corruption of test taker output in OPIs was found by Reemann, Alas, and Liiv (2013), who discovered that deviations from the OPI script by interlocutors correlated with their gender in Estonia. In their study, female interlocutors tended to be more accommodative and generous with time, which resulted in lower test reliability. This perhaps demonstrates that despite present-day procedures of standardization and other strict interlocutor procedural guidelines in fixing input frames, a human element in test delivery will always include the possibility of error or aberration. And conversely, test takers in OPIs, through the presence of a human interlocutor, are afforded the opportunity to interrupt the relatively formatted test input discourse, which is an additional nuance to test reliability.

Arguments supporting the continued use of OPIs are partly based upon the notion that they can provide opportunities to elicit a wider range of discourse moves from test takers and therefore it is claimed that they could more accurately reflect spoken interaction. Seedhouse and Morales (2017) have proposed a fourth part to be included into the IELTS OPI, enabling interlocutor question followed by test taker response discourse exchanges to be reversed. Results from their study suggest that these atypical sequence of exchanges were successfully elicited, along with a variety of other speech acts. Furthermore, additional rating data were retrieved from the performances throughout this proposed new section that provided a clear distinction between stronger and weaker test takers. Nonetheless, introducing candidate-led interaction into OPIs may affect opportunities for test takers of varying ability to demonstrate their competence and thus affect the reliability of the test.

Questions as to the reliability of SOPIs have been put forth by Young-Ju (2007) which also affect CB-OPIs. Namely, Young-Ju (2007) discovered that SOPI raters tended to listen to recordings selectively. For instance, raters would begin listening to task three first, and if the test taker performed well there, would then move to task four or five. However, they would skip these subsequent tasks and focus on tasks one or two if the performance in task three was not successful. Selective listening deprives the examinee of a holistic appraisal of their interview and is important due to the frequently given expectation that test taker performance is rated throughout all parts of a test. This example of selective listening to test taker performance by raters cannot be as easily amended in OPIs. Young-Ju's (2007) findings

were based on a sample size that was too small to determine if this selective listening behavior affected reliability; nevertheless, this does warrant further investigation. It is worth noting, however, that the irregularity identified by Young-Ju (2007) is not associated with the delivery of the test and hence could be rectified during the rating process.

During the documented recent revision of the *Aptis* CB-OPI speaking rating scales, Fairburn and Dunlea (2017) noted that although inter-rater agreement was high using the new scales, minor inconsistencies in the rating of performances were primarily attributable to the rater effects of leniency or severity and a tendency to mark centrally within the scale producing a restricted range of scores. Considering these findings that well-developed rating scales and rater training and standardization can improve the accuracy and consistency of test scores, much is still unknown of the cognitive processes employed by raters when making scoring decisions and further research investigating this construct irrelevant facet would be beneficial.

One way of increasing the reliability of test scores is to utilize double-marking of test event performances by raters. Through the use of pre-marked control items regularly interspersed within the performances to be rated, both inter- and intra-rating can be monitored. Whilst the benefits of incorporating second-marking are numerous, questions remain as to its accuracy for OPIs, because a live test event performance cannot be replicated. The use of audio and audio-visual recordings of OPIs for double-marking has recently been brought into focus with Beltran (2016) reporting no significant differences between the rating of audio and audio-visual speaking test performances, because despite there being variability between scores, these remained within a comparable range. However, the raters participating in this study expressed a preference for the audio-visual input mode, due to it "allowing for a more complete and straightforward scoring experience" (Beltran, 2016, p.15).

Nevertheless, a smaller study investigating the rating of live, audio-visual and audio only input modes of the IELTS speaking test conducted by Nakatsuhara, Inoue and Taylor (2017b) reported significant differences in the scores awarded for audio recorded spoken performances and live test events. It was proposed that this finding demonstrates narrower constructs are being assessed in the audio input mode. Consistency between scores of performances in the audio-visual and live modes was reported as being much closer as raters were able to notice paralinguistic features that contribute toward discourse expected in *real-life* interaction.

Semi-direct speaking test score consistency has been shown to possess the reliability which is required to facilitate high-quality decisions. This is important as interpretations must be meaningful, impartial, generalizable and relevant. Furthermore, the stability of interlocutor input throughout successive administrations of a test along with the degree of rater agreement achievable when a test event performance is scored by audio only suggests that reliability is a test quality wherein semi-direct tests can claim interchangeability with their direct counterparts. However, concurrent validity evidence of test scores between these two delivery modes that demonstrate the predictive ability of CB-OPIs to accurately reference performance, and thus scores, in OPIs is also essential.

CONCURRENT VALIDITY

As it is proposed that semi-direct speaking tests be substituted for OPIs, concurrent validity is of high importance to all test stakeholders. Much of the research investigating direct and semi-direct speaking test interchangeability has focused on concurrent validation that involves equating test scores between

direct and semi-direct tests. This research has often used the OPI as the criterion behavior, with the presumption that it is always the semi-direct tests' validity that needs to be established. Clark (1979) initially argued that a correlation figure of 0.9 or higher would be an appropriate level of agreement to suggest interchangeability; yet a subsequent study by Clark and Swinton (1980) found that scores on the TSE and OPI correlated at $r = 0.80$ and they concluded that the TSE was a reasonable alternative.

With the Center for Applied Linguistics' introduction of the first SOPIs designed to test less commonly taught languages in the late 1980s, subsequent research from Stansfield (1991) reported Pearson correlations of 0.89 – 0.95 in the OPI and SOPI variants leading to the proposal that the SOPI correlates so highly with the OPI that it seems safe to say that both measures test the same abilities. Moreover, on the basis of Stansfield and Kenyon's (1992a, 1992b) research, it was concluded that the OPI and SOPI are highly comparable measures of oral language proficiency and can be viewed as parallel tests in two different formats.

More recent studies investigating the concurrent validity of CB-OPIs include those by Rosenfeld, Bernstein, and Balogh (2005) and Mousavi (2009) and have found that comparisons of test taker scores between the delivery modes correlate highly, signaling the possibility that no adverse effects are to be found. Additionally, Surface, Poncheri, and Bhavsar (2008) conducted an OPI to CB-OPI concurrent validity study with 99 Korean learners of English and found that absolute agreement on final scores between the two tests was 63%, but increased to 98% when considering adjacent agreement alongside the absolute rating. Moreover, Thompson, Cox, and Knapp (2016) found that 55% of test-takers who took both an OPI and CB-OPI had the same score on both exams with this percentage increasing to 97% for scores with adjacent category agreement.

The most recent published study, Dunlea et al. (2018), had very similar results to both the Surface et al. (2008) and Thompson et al. (2016) studies, but used a test that measured all four skills. Comparing 384 Vietnamese students who took both a paper based English proficiency test and the *Aptis* CB-OPI, Dunlea el al. (2018) found 66% exact agreement, and 99% agreement when including adjacent placement. From the results of the above-mentioned studies, it can be seen that exact agreement typically falls below acceptable standards, yet is impressively high when accounting for adjacent agreement. Adjacent agreement could be a fair concession in justifying the overall validity between direct and semi-direct OPIs; however, serious repercussions within the allowance could exist. For instance, if adjacent scores straddled threshold levels, such as intermediate-high and advanced-low, a test taker may be affected by one test format over the other if entry requirements sought only advanced placements.

Yonezaki (2016) conducted a comparison study between direct and semi-direct speaking assessments with Japanese participants and found no statistically significant differences between scores. This study also compared the speaking criteria separately to determine if specific skills may differ between test formats. Although fluency and accuracy were slightly better in the direct test, and the other two criteria of volume and content showed marginally superior numbers for the semi-direct test, Yonezaki (2016) concluded no significant differences among these scoring criteria between test formats.

Kim and Craig's (2012) interesting experimental study validating a CM video conferenced speaking test compared this delivery mode with a face-to-face OPI and found "no significant difference in performance between test modes, neither overall nor across analytic scoring features" (p.257). In addition, Ockey, Koyama, Setoguchi, and Sun (2015) compared the *TOEFL iBT* speaking scores of 222 Japanese university students to their performances on face-to-face or group oral assessments in their university English classes and also found a strong relationship between their performances.

Kim and Craig's (2012) validation study results were later found to be representative as Nakatsuhara et al.'s (2017b) comparison between video conferenced OPIs and face-to-face OPIs also found no significant differences in overall test scores or in performance in rating scale criteria (fluency, lexis, grammar, and pronunciation). Nakatsuhara et al. (2017b) also analyzed language functions elicited from each test format and from 30 language functions examined, only clarification, comparison, and suggestion showed significant differences in use between test modes. The differences were explained by questionnaire feedback which showed that the sound quality of the video conferencing tests was moderately unreliable in part one of the test, where the differences were found.

As evidenced from the research presented, there is some degree of concurrent validity between direct and semi-direct speaking tests, but Shohamy (1994), Weir (2005) and Zhou (2015) agree that high correlations between delivery modes of a test provide necessary but insufficient evidence of their interchangeability, because of the need to investigate the equivalence of direct and semi-direct speaking tests from multiple perspectives. Concerns generated include lower test validity because of the possibility of obscuring or introducing mode-specific features indicating potential disparities that may prove detrimental to the way in which assessment scores are interpreted and include test taker spoken output.

TEST TAKER OUTPUT

Early Speaking test theorem argued that direct tests are the better measures of speaking ability due to their close relationship between test context and *real-life*, yet acknowledged that the language elicited is unrepresentative of *real-life* conversational discourse, because test takers are aware that they are talking to a language assessor (Clark, 1979).

A relatively small number of studies have investigated the equivalence of test taker output in direct and semi-direct speaking tests, and those which have (Choi, 2014; O'Loughlin, 2001; Shohamy 1994; Zhou, 2008) have used closely matched tasks, serving to complicate potential output divisions due to the difficulties experienced in separating task effect from mode effect. The following abridged version of a recent published study entitled *Output register parallelism in an identical direct and semi-direct speaking test: A case study* (Quaid, 2018) represents an example of research investigating this facet of test mode comparability. For brevity, the introduction and conclusion sections have been excluded from presentation, whereupon the study can be joined at the following literature review stage that details previous research findings of salient language register features to be investigated in the study.

LITERATURE REVIEW

Rhetorical Functions and Structure

Shohamy (1994) defined rhetorical functions as the nature of the input or prompt used to elicit test taker output, while rhetorical structure referred to the underlying discourse organization of the speech event in its entirety. Shohamy (1994) found both were different in her study, because the OPI mainly consisted of direct interrogatives and the SOPI of declarative instructional prompts. Two possible explanations for these differences could be task effect or the now relatively dated speaking test used in the comparison, which predominantly used a *task-response-new task* rhetorical structure. The latter is

a more likely explanation as differing types of interaction have been included in more recent CB-OPIs that allow inferences to be made about performance in a myriad of contexts (Laborda, 2010), and thus later studies (O'Loughlin, 2001; Choi, 2014) have shown that the rhetorical functions and structure are highly comparable in monologic tasks of direct and semi-direct speaking tests.

Agreement of research findings relating to the similarities between the rhetorical functions and structure in test taker output in direct and semi-direct versions of monologic test tasks appears to be within sight. However, less agreement has been reached with regard to the similarity of prosodic features and contextualization in test taker output between delivery modes.

Prosodic Features and Contextualization

Shohamy (1994) and O'Loughlin (2001) agreed that the prosodic features of test taker output differed, with the semi-direct tests monologic tasks encouraging less variation in intonation and voice range: Kiddle and Kormos (2011) concurred with Shohamy (1994) and O'Loughlin's (2001) earlier research findings in their study, which compared test taker output from an OPI with a CB-OPI using a purpose developed speaking test. Results showed that there was a statistically significant difference in pronunciation performance in monologic speaking tasks, when the test taker response was delivered via a microphone on a computer, yet in a more recent study, Choi (2014) noted that the OPI and SOPI performance samples did not show any distinctive patterns or characteristics in terms of average tone and pitch shifts, and thus concluded that prosodic features were very similar.

Shohamy (1994) reported on the differing test taker spoken discourses elicited by the two versions of the test and contextualization was proposed as a major contributing factor. This was signaled by test taker laughter, addressing of the interlocutor directly, sharing personal information and referring to physical objects located within the testing environment to explain or demonstrate a point being made. Shohamy (1994) remarked that "context alone seems to be more powerful than the elicitation tasks themselves" (p.118). Nonetheless, Shohamy, Donitsa-Schmidt and Waizer (1993) did concede that it was unclear if these facets of contextualization were a result of not controlling for task effect. However, O'Loughlin (2001) found no evidence of similar contextualization in either version of the *access* speaking test. Furthermore, Choi (2014) labelled contextualization as being very similar in both delivery modes.

To examine contextualization is one method of determining interpersonal differences in elicited test taker discourse; however, lexical density analyses have provided an alternative tool with which to investigate possible output register variation.

Lexical Density

Shohamy (1994) found that the OPI output contained 40 per cent lexical items and 60 per cent functional items approximately whilst these figures were reversed for the semi-direct test. On this basis, it was argued that the language output possessed greater literateness in the semi-direct test. Similarly, O'Loughlin (2001) reported a smaller but statistically significant difference between modes, and suggested that the lack of interlocutor verbal feedback during the semi-direct test could explain this difference, because even in monologic tasks during the direct test, interlocutors provided reactive tokens to the test takers.

Zhou (2008) found her lexical density results conflicted with those of Shohamy (1994) and O'Loughlin (2001), because there were no significant differences between the OPI and CB-OPI. Zhou (2008) proposed the disparity of this finding "may be attributable to that the tasks used… were more parallel in various

aspects across modes" (p. 203). It is likely that Shohamy (1994) and O'Loughlin's (2001) inclusion of open tasks was a key contributor to the task effect noted in their studies, as ultimately the range and frequency of lexical items within test taker output are subject to the demands of the task.

Choi (2014) documented that output from the semi-direct test had a slightly higher lexical density than the OPI throughout all task types in his study. Whilst agreeing that task type contributed to output lexical density, Choi put forward an alternate proposition to O'Loughlin (2001) in that his closed narration tasks elicited more lexically dense output than the open description tasks. This discrepancy may well be due to the two studies not being comparable, because the majority of Choi's (2014) tasks allowed for test taker planning time, which could be a source of lexical density variation.

Syntactic Complexity

Shohamy (1994) reported a possible difference between connecters used in the OPI and SOPI. A subsequent study conducted by Luoma (1997) agreed and showed that there was a possible difference in a tendency towards more coordination on the tape-mediated test. Luoma (1997) showed that there were almost twice as many coordinating conjunctions during the performances in the tape-based test, but a ratio of only 4:3 respectively in test taker output from the face-to-face test. The sets' range of coordinating and subordinating conjunctions were almost identical between modes, although some less frequently used were perhaps more mode specific, and may have implied slight register variation. During the face-to-face test, informal vague list completers such as *and_stuff* and *and_everything* were found, yet none were discovered in output from the tape-mediated test.

Luoma (1997) also noted unexpected de-contextualization of the direct test event, which was found within subordinating conjunction use in addressing the interlocutor, whereby test takers used *as_it_says* during the face-to-face test, whilst deciding to acknowledge the speaker in the tape-mediated equivalent with *as_you_have_said*. Shohamy (1994) found that the OPI test taker output contained a variety of discourse connectives, whereas these were restricted during the SOPI. In direct contrast, O'Loughlin (2001) reported that output from both versions of the *access* test's monologic tasks showed an equally broad range of discourse linkers, but that they were used less frequently during the semi-direct test.

Choi (2014) found that the types and numbers of discourse connectives and markers in test taker output were identical in both the direct and semi-direct versions of the test tasks used in his study. These findings are interesting because we may expect a number of specific cohesive devices to be found in test taker output from semi-direct speaking tests, illustrating what many researchers (O'Loughlin, 2001; Shohamy 1994; Zhou 2008) have described as a more formal register or style, with a higher degree of literateness and cohesion.

From syntactic complexity analyses of test taker output between delivery modes at AS-unit and clausal level, while using monologic tasks, Zhou (2008) reported no significant difference in the syntactic complexity of the output examined, although the means of syntactic complexity measures were slightly higher in the CB-OPI. Furthermore, Choi's (2014) study, also using monologic tasks, similarly showed that the semi-direct test was prone to elicit longer and more complex utterances.

RESEARCH QUESTIONS

1. Can test taker output register equivalence be shown in an identical semi-direct and direct version of the *Aptis General* speaking test?
2. What output features are primary contributors toward any register shift found?

METHODS

The present study used an alternate methodological approach to determining test taker output equivalence in direct and semi-direct speaking tests. In contrast to all previous studies referenced in this paper, identical test task type and content were used and interlocutor effect was successfully minimized. Figure 1 provides an overview of the counter-balanced study design.

MATERIAL

Aptis General CB-OPI

The *Aptis General* Speaking test (British Council, 2018) was used to form the basis of this study. The test is delivered as a CB-OPI and consists of four parts using monologic tasks, whose full specifications are available in O'Sullivan and Dunlea (2015). A number of these specifications are identical throughout all four tasks, such as the CB delivery mode and the pre-recorded and scripted features of the tasks, but task feature specifications that do differ by task type and/or number are summarized below.

The online practice version of the *Aptis General* speaking test (2016) was chosen in preference to current live versions to minimize test security risks, enable immediate access to data for stimulated recall

Figure 1. Study design

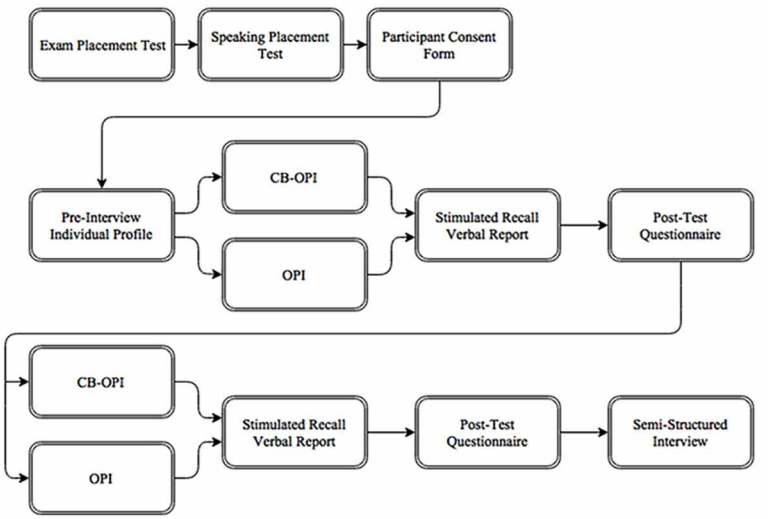

Table 1. Task feature differences

Specification	*Task 1*	*Task 2*	*Task 3*	*Task 4*
Task level (CEFR)	A1, A2	B1	B1	B2
Rubric presentation	Aural, written	Aural, written, visual	Aural, written, visual	Aural, written, visual
Response format	Q&A	Q&A	Q&A	Long turn
Interaction	Interlocutor - candidate	Interlocutor - candidate	Interlocutor - candidate	Candidate only

verbal report interviews and allow for unedited publication of this study. No changes were made to this standard version of the *Aptis* General speaking test for delivery. The CB *Aptis General* speaking test was fully transcribed, interlocutor input modified and visual prompts printed for face-to-face delivery.

Participants

A pool of 30 participants was invited to complete Cambridge English Language Assessment's (2016) Adult Learners' General English exam placement test (EPT). Four respondents were identified from their EPT scores as minimally lower-intermediate English language users and completed a face-to-face speaking placement test (SPT) in order to establish the validity of the scores attained from the EPT with regard to their speaking ability. Pseudonyms were assigned to all participants for the reporting of this study to protect identities and maintain confidentiality agreements.

The participants were female and of Chinese nationality with a corresponding L1, who were 18-35 years old and had excellent familiarity with computing and technology. Emily and Katrina were studying English in company-organized classes at the time of the data collection sessions, but only one lesson was completed during the five-day span of the data collection phases, and this related to salutations and closings in business emails and therefore was not deemed to be of consequence for this study's results. All four participants reported some familiarity with the *Aptis General* Speaking test format, although none had completed the online practice version of the test used in this study.

Table 2. Participants' EPT score, recommended exam and SPT Level

Participant	*EPT score*	*Recommended Cambridge English exam*	*SPT level*
Elena	22	Cambridge English Advanced (CAE) / Cambridge English Proficiency (CPE)	Upper-intermediate
Chloe	25	Cambridge English Proficiency (CPE)	Intermediate
Emily	24	Cambridge English Proficiency (CPE)	Intermediate
Katrina	21	Cambridge English First (FCE) / Cambridge English Advanced (CAE)	Lower-intermediate

Procedure

The participants were divided into two groups that were assigned to complete alternate test delivery formats during each of the two data collection sessions. The CB practice version of the *Aptis General* Speaking test was administered in standard format in a computer laboratory, and participants' spoken performances were captured by an audio DRD recorder. The face-to-face OPI delivery was conducted in a quiet room adjoining the computer laboratory.

Situational practicality demanded that the two data collection sessions were conducted within a five-day period. No unreasonable practice or recency effect were noted during the second data collection session. Participants were not informed of task nature or content before either of the sessions. The instructions to participants and researcher protocol for the stimulated recall interviews were adapted from Mackay and Gass (2000). The self-report interviews were conducted immediately after each administration of the speaking test. The DRD recorder containing the participants spoken output from the OPIs was placed between the participant and the researcher, and both parties were able to pause the recording on listening to the tests to comment or ask questions. A limitation of selecting the online practice version was that the test was unable to be paused after each task for the verbal report to be conducted using the audio recordings as stimuli, and therefore the full tests needed to be administered before initiation of the stimulated recall phases.

The second data collection session differed by inclusion of a semi-structured interview conducted at the end of the session that allowed participants to expand on and contrast their responses in the two post-test questionnaires. This interview was co-constructed in that both the participant and the interviewer could select salient responses from the post-test questionnaires to form part of the interview.

Analyses and Results

Spoken output data from the four participants were transcribed following conversation analysis transcription notation conventions that were recommended by Atkinson and Heritage (1984), and subsequently adopted by Lazaraton (2002), to investigate test taker output in OPIs. Transcription of the spoken performances was primarily carried out by the lead investigator; however, twenty-five per cent of the raw data was double-coded by a trained research assistant to validate the accuracy of the coded transcripts. The spoken data from the stimulated recall and semi-structured post-test interviews were not transcribed, save for excerpts presented in this study.

Foster, Tonkyn, and Wigglesworth (2000) define AS-units as being "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (p. 365). Foster et al.'s AS-unit was chosen as the basic syntactic speech element best suited to the syntactic complexity analyses in this study, as it was considered a better measure than others, such as T- and C-units which are inadequate to deal with a full analysis of spoken discourse, because multiple unit definitions inhibit studies being replicated with any real accuracy.

Rhetorical Functions and Structure

The rhetorical functions and structure of the performance samples elicited from the OPI and the CB-OPI were comparable. The variance of item input in the test, containing both task instructions and direct questions, ensured maximum flexibility of the discourse event. Non-conformation to standard *question–*

answer–question traditional OPI form, or *task-performance–new task* discourse units, typical of some early semi-direct speaking tests, encouraged a wider range of discourse exchanges: *task–performance-question–answer*. Deviation from this myriad of exchange forms was rare, and was only found in the OPI when test takers made requests for instruction clarification:

[Katrina: OPI, Task 4 Instructions]

116 IN you now have two minutes? to talk?

117 PD for each question or all the three? (.2) the three ok. .hhh! (.)

((looks at interlocutor who nods))

Test taker output was similar between modes at a macro-structure level, sharing many common features. The following excerpts clearly illustrate the similar discourse organization of the texts elicited:

[Elena: CB-OPI, Task 1, Item 2]

17 EL well basically? there is no pattern or theme for me t(hhh)o do in

18 my free time, .hhh! (.2) um i just simply like (.) reading books.

*19 (.) (*er watching*) some films (.2) or listen to some music, .hhh*

*20 (.5) and it really all depends on my (*moods,*) (.2) <not*

21 some> kind of thing i have to do. .hhh (.5) so basically free

*22 time (*for me*) is, do >whatever i want?!< (*and*) do it*

*23 <whenever (*i want?*)> (.5) ok,*

[Elena: OPI, Task 1, Item 2]

19 EL (%ah?%) well what i like to do in my free time is to. (.) do.

20 almost (.) nothing, or just lying down? .hhh ah reading some

21 books? or watching some tvs, .hhh (.) so? basically free time is

22 for me is to do >whatever i want,< there's no such thing

*23 as, .hhh (.) (*um*) tch like i'll go? (.) jogging or something*

*24 like >that it?< all depends on my (*mood,*) (1.5)*

25 (%ok?%) (.) ah (%i'm ready.%) tch

The discourse functions in test taker output at the level of act were also highly comparable throughout the tests.

Contextualization and Prosodic Features

A higher degree of contextualization was found in the output from the direct OPI. Discourse markers indicative of more interactive exchanges were found throughout test taker output in the direct versions of the test, and their relative absence within the elicited discourse of the semi-direct CB-OPI was marked. A salient example was Elena's frequent use of *you know*, when attempting to seek interlocutor agreement:

[Elena: OPI, Part 3, Item 1, Response]

82 playing basketball you can see a lot of people,.hhh (.) you know?

*83 chasing the balls? (.) (*and*) you know fighting er with*

During Elena's output in the CB-OPI, no seemingly direct attempts were made to address the interlocutor. The second person pronoun *you* was instead used in isolation to achieve personal reference to a lesser degree. To exemplify this, the following excerpt from CB-OPI output sought to advise personally and generically of entry to a museum:

[Elena: CB-OPI, Part 2, Item 2, Response]

*52 (*er*) it is free admission so you can go >whenever< you want*

This both generic and personal use of *you* by participants was frequently observed in both test delivery modes, and individual test task items were almost certainly an important factor in its use.

Discourse particles were used frequently by all participants in the OPI at the end of turns, to seek agreement from the interlocutor of the quality or length of their response. The most frequent of these was *OK?* used with rising intonation:

[Chloe: OPI, Part 3, Item 3, Response]

92 so (.2) um: basketball is more difficult. it's (.8) ok?%

This directly contrasted with their use in the CB-OPI as a hesitation filler at the beginning of turns or as a signal of turn completion, which were equally identifiable by the use of flat continuing intonation:

[Katrina, CB-OPI, Part 1, Item 1, Response]

*10 PD ok. (.) (*there a:re four*) members in my family. (.) .hhh! my*

Additional evidence of test taker contextualization during the OPIs was found in laughter, requests for clarification of task instructions and in more interpersonal turn completion signals used with rising intonation.

Lexical Density

To give an overall impression of the performances in each mode, word counts were taken per task and item for each participant. A word count calculation resulted in overall test means per delivery mode of CB-OPI *782.75:797.75* OPI. This result, coupled with the individual task/item word counts in Table 3, suggested that both versions elicited a very similar amount of ratable language.

To investigate claims of spoken output register variation between delivery modes, lexical density analyses of the test takers' performances were conducted. The taxonomy framework originally proposed by Halliday (1989), and subsequently revised and used by O'Loughlin (2001) and Choi (2014), was adopted as the item classification system used in these analyses. It was decided that no distinction be made between high and low frequency lexical items, so as to be able to effectively compare results with the most similar and recent studies (Choi 2014, Zhou 2008), as well as those which were broadly relevant to this current investigation (O'Loughlin, 2001; Shohamy, 1994).

An analysis of all collected spoken data produced very similar overall lexical density mode percentage medians of CB-OPI *40.614:40.329* OPI. Figure 2 shows these overall mean values distributed into percentage figures per test task item.

The CB-OPI elicited minimally greater lexically dense output in two task items, although more noticeable is a strikingly similar overall pattern in the lexical density of output per individual test task item between delivery modes.

Syntactic Complexity

An analysis of the total occurrence of all conjunctions was found to be CB-OPI *231:277* OPI. Therefore, a relatively even distribution was found between modes. A similar ratio of coordination and subordination was also present: CB-OPI coordination *65.4:34.6* subordination, and OPI coordination *67.1:32.9* subordination. Table 4 presents occurrence by conjunction type, identifier, number used, and their frequency of use among all conjunctions found in the tests.

To determine complexity at the level of AS-unit, three additional analyses were conducted and the first of these was the number of words per AS-unit.

Table 3. Mean output word count per task item

Delivery mode	*Task 1*			*Task 2*			*Task 3*			*Task 4*
	Item 1	*Item 2*	*Item 3*	*Item 1*	*Item 2*	*Item 3*	*Item 1*	*Item 2*	*Item 3*	*Item*
CB-OPI	50.5	49	48.25	72.5	78	80.75	70.75	65.5	74.25	193.25
OPI	53.75	56	46.75	74.75	76	72.75	69	65.25	89.25	195

Figure 2. Mean lexical density percentage values per test task item

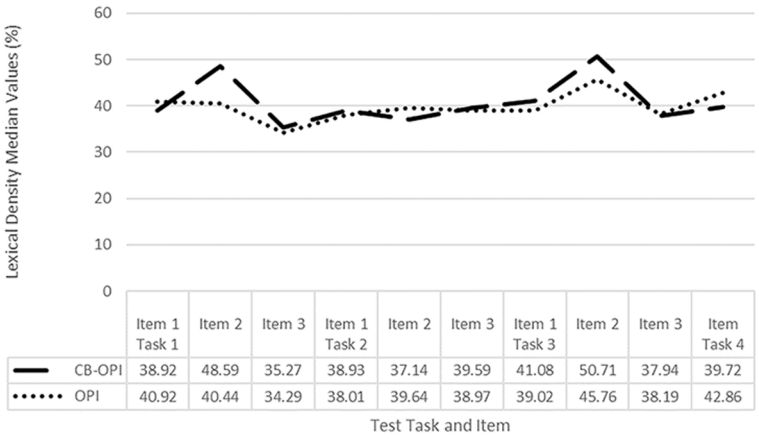

Table 4. Phrasal and clausal conjunction type and frequency

CB-OPI coordinators			*OPI coordinators*		
Conjunction	*No.*	*%**	*Conjunction*	*No.*	*%**
and	100	43.2	and	109	39.3
so	19	8.2	so	35	12.6
but	18	7.7	but	21	7.5
or	14	6	or	21	7.5
not only… also	1	0.4			
CB-OPI subordinators			*OPI subordinators*		
Conjunction	*No.*	*%**	*Conjunction*	*No.*	*%**
because	30	12.9	that	24	8.6
that	19	8.2	because	21	7.5
when	8	3.4	when	16	5.7
which	4	1.7	if (whether)	9	3.2
as	3	1.2	what	6	2.1
if, how, what, whatever, while	2 each	0.8	whatever	3	1
			even though, how	2	0.7
			however, who, as, which, after whenever, while, as long as	1 each	0.3
since, whenever who, where, without,	1 each	0.4			

* Percentage of conjunction in test mode for ease of comparison

Three participants used more words per AS-unit during the CB-OPI; however, Elena's output did not conform to this overall trend, because she used more words per AS-unit in the OPI. The second analysis investigated the number of clauses per AS-unit, and within these results a pattern for all participants emerged. Table 6 shows a greater percentage of clauses per AS-unit in output from the CB-OPI, albeit marginally for Emily and Katrina.

Finally, an analysis of the number of subordinate clauses per AS-unit revealed that three of the four participants used moderately more subordinating clauses in the CB-OPI, whilst Elena used subordination equally in both versions of the test.

The results of the syntactic complexity analyses were comparable between modes. The OPI elicited minimally more linking of phrases in the output retrieved, but at clausal and AS-unit level the CB-OPI output possessed only minimally levels of higher syntactic complexity and subordination.

Discussion

The rhetorical functions and structure of the discourse during the test events were very similar and this was perhaps most relevant to the test itself, because findings between studies have differed according to the nature of the task item input. Atypical from the early version of a semi-direct test that formed the basis of Shohamy's (1994) study, a combination of questions and task prompts are used to form input in the newer *Aptis General* CB-OPI. This results in a relatively more varied speaking test discourse in both direct and semi-direct delivery formats.

Prosodic feature differences were not immediately apparent when analyzing the transcribed and fully-coded test performances. However, on re-listening to the complete audio-recorded tests to gain an overall impression, it was difficult not to agree with previous research findings (Kiddle and Kormos,

Table 5. Words per AS-unit

Participant	*Elena*	*Chloe*	*Emily*	*Katrina*
CB-OPI	10.46	11.22	10.69	10.88
OPI	11.02	10.08	10.55	10.13

Table 6. Clauses per AS-unit

Participant	*Elena*	*Chloe*	*Emily*	*Katrina*
CB-OPI	1.64	1.64	1.52	1.51
OPI	1.54	1.55	1.48	1.49

Table 7. Subordinate clauses per AS-unit

Participant	*Elena*	*Chloe*	*Emily*	*Katrina*
CB-OPI	0.47	0.57	0.4	0.3
OPI	0.47	0.43	0.36	0.2

2011; O'Loughlin, 2001; Shohamy, 1994) that monologic-tasked semi-direct tests encourage marginally less variation in intonation and voice range and sound more monotonous.

A more concrete difference was that in the OPIs paralinguistic and prosodic features were often responsible for contextualization by way of test takers attempting to build a relationship with the interlocutor. For example, and exemplifying the non-verbalized contextualization frequently found in the OPIs, Elena reported that she glanced at the interlocutor, who was checking the timer, when she suspected her response was too short in part three. From this non-verbal exchange, she interpreted the need to continue her turn. Furthermore, a verbalized example found more frequently was Elena's use of *you know* when attempting to seek interlocutor agreement. It was unclear from the spoken data from the OPI whether this was expressed consciously or unconsciously and if it was rhetorical in nature or if some means of non-verbal response was desired.

The use and pronunciation of discourse markers for contextualization was prevalent throughout the OPI performances and were normally marked by rising intonation. A possible explanation for the test takers interpersonal behavior was revealed in the post-test interviews, when Katrina explained that it "felt like she was actually talking to someone", and Elena said that "it really felt like an interaction between two people". This insinuates that the mere presence of an unresponsive human interlocutor can affect the tenor and register of test taker output and that contextualization and interactiveness may well have been the most sensitive measure of difference in this study, although this degree of interactiveness had a relatively minor effect on the quantity or lexical density of the elicited spoken output.

Along with the findings of previous studies (Luoma, 1997; Shohamy, 1994; Zhou, 2008), overall output word counts were similar between modes. In general, the CB-OPI elicited minimally higher lexically dense output, supporting previous related research (Choi, 2014; O'Loughlin, 2001; Zhou, 2008), which partly controlled for interlocutor effect and used closely matched or identical monologic test tasks. These results imply that Shohamy's (1994) findings of significantly greater lexical density in the semi-direct test may not be robust and may have resulted from task and interlocutor effect.

Lexical density variance was found between individual test task items. Conforming to lexical density theory, less literateness was generally demonstrated during test takers' response to the everyday conversational and contextualized input; whereas more was found in output elicited from decontextualized situational items.

Phrase level syntactic complexity was relatively similar in test taker output from the direct and semi-direct test in this study, although slightly more conjunctions overall were found in the output from the OPI than the corresponding CB-OPI. This disagreed with Luoma's (1997) earlier finding that the overall difference in frequency of conjunction use between modes was much larger. A possible explanation for this greater difference may be that the test tasks used by Luoma were not identical in structure.

Luoma (1997) showed that there were almost twice as many coordinating conjunctions during the performances on the tape-based test, but a ratio of only 4:3 respectively in test taker output from the face-to-face test. This finding was also not replicated in this present study, because although performances in both delivery modes contained a higher frequency of coordinating conjunction use, the ratios between coordination and subordination conjunctions were relatively stable between modes. However, the supposition of minimally more complex grammatical structure being representative of test takers' output in CB-OPIs was confirmed, when examining the participants' performances at AS-unit and clausal level.

Overall, more words per AS-unit were used in the CB-OPI, and differences were also found in the number of clauses and subordinating clauses per AS-unit, with the CB-OPI output containing minimally higher syntactic complexity means. These findings corroborated those of Choi (2014) and Zhou (2008),

although any difference in syntactic complexity between modes was unlikely to be significant should larger test taker output data samples have been analyzed. Nonetheless, Brooks and Swain (2014) have shown a significant decrease in syntactic complexity in output from the speaking section of the CB *TOEFL iBT™* to conversation in university contexts. Syntactic complexity is therefore likely to increase as spoken discourse becomes of a less interactive and formal nature.

Nevertheless, the importance of the syntactic complexity results for this present study should not be underestimated, because the performance samples elicited in the CB-OPI were "more likely to contain longer and more complex sentences" (Choi, 2014, p.563). Responsibility for this can only be reasonably assigned to mode effect, inside which interlocutor effect and the resulting increased test discourse interactiveness are intricately and unavoidably interwoven.

Summary

Test taker output register equivalence is perhaps unachievable in identical semi- direct and direct versions of speaking tests, as there may be the tendency for CB-OPIs to elicit test taker output leaning toward the minimally more formal, written and literate. This register variance is primarily attributable to interlocutor effect, a facet of mode effect, which leads to a higher degree of interactiveness in direct test events.

However, the tendency for semi-direct tests to elicit minimally higher lexical density and syntactic complexity in spoken output should be seen as advantageous for both test taker and delivery mode. Lexical and grammatical range and complexity are frequently found in band descriptors of criterion-referenced speaking tests, and in encouraging the use of more lexically dense and syntactically complex output, semi-direct tests more accurately reflect true ability.

Case Study Limitations

Several limitations should be remembered when interpreting the results of this case study. Firstly, the study was necessarily sample dependent, which prevent the generalizability of findings because the participants were from a narrowly defined population that is unlikely to be fully representative of a broader second language English user population. Secondly, the use of identical task type and content could have encouraged recency effect in the participants' performances during the second administration of the test. And the possibility of increasing familiarity with the interlocutor was also present, as the researcher acted as both interviewer and interlocutor throughout the sampling and data collection phases of the study. It is also feasible that if a female had acted as the interlocutor, performance results may have differed (O'Sullivan, 2000).

FUTURE RESEARCH DIRECTIONS

It is clear from the four selected test facets synopsized in the first part of this chapter that evidence through progressive studies has provided a foundation on which to continue to develop and implement semi-direct speaking tests. The use of CM video-conferencing in second language speaking assessment is currently an under researched delivery mode and has been less widely used than CB testing. Further research on the CM video-conferencing test delivery mode's practicality could focus on test takers' untrue perception that sound quality affected their performance when, in fact, the connection remained stable

(Nakatsuhara et al., 2017a). Similarly, further investigations into the online stability of CB tests input and written and/or visual prompts are essential if inclusiveness, via equal opportunity and fairness, for test takers is a primary aim of semi-direct speaking test delivery.

Much research has been conducted on test taker affective variables toward direct and semi-direct speaking tests. Reduced nervousness during completion of a semi-direct speaking test has been experienced by many, due to a human interlocutor not being present. This is an area of investigation that could support the face validity of CB speaking tests with further generalizable research. Additional studies are also needed to examine if test takers feel that CB speaking tests accurately measure the constructs they purport to.

CB-OPI test-retest reliability studies have been generally positive, although additional themed research as new tests are developed would further support findings to date. In terms of rater reliability more research is needed to explore the cognitive processes employed when completing the marking of audio and audio-visual speaking tests, including the rater effects of leniency or severity and a tendency to mark centrally within rating scales producing a restricted range of scores. In addition, studies addressing how the gender of interlocutors affects test taker performance and the test event more generally would be beneficial.

The concurrent validity of test scores between delivery modes has been found to be highly accordant when accounting for adjacent agreement, yet when this is excluded, exact agreement can fall below acceptable tolerances. Although adjacent agreement of scores may be acceptable for low stakes tests, it is less so for those of high stakes denomination. It is thus important for these tests to regularly commit to improving absolute agreement through ongoing research.

The exampled case study in the latter part of this chapter suggests that future research could include a larger generalizable comparative replication study or an investigation of interlocutor effect's relationship with output register, by administering OPIs to test takers with unfamiliar and moderately familiar interlocutors and examine the contextualization and interactiveness in the spoken output retrieved. An awareness of how familiarity in the interlocutor–test taker relationship can affect output in OPIs could influence speaking test delivery practices.

CONCLUSION

The purpose of this chapter was to examine the equivalence of semi-direct CB and CM speaking tests and their more traditional face-to-face direct counterparts through four essential test facets, indicative of their interchangeability, and highlight the comparability of test taker output elicited through direct and semi-direct input delivery modes.

The practicality advantages of CB and CM speaking tests are marked for test takers, test score users and large-scale test developers, yet for smaller would be institutional level developers, costs can be prohibitive. Through a number of studies investigating the face validity of direct and semi-direct speaking test delivery modes, it has been found that test takers' affective behavior toward direct OPIs often acknowledges that a perceived, but untrue, fundamental change takes place in that they are engaged in a *real-life* conversation during the face-to-face test. It is therefore likely that direct speaking tests' current higher face validity is a primary justification for their continued use.

Numerous benefits are apparent for the development and use of CB-OPIs with regard to the reliability of input achievable and test scores. The absence of a human interlocutor ensures that task input is stable

within and between test events and thus the factor of deviation in the resulting spoken discourse is a non-entity. Perhaps the most significant challenge for CB and CM delivery modes is the effective rating of test takers' audio-visual and audio only output, given paralinguistic features indicative of spoken discourse are absent.

To date, concurrent validity has been one of the most highly researched areas of semi-direct and direct speaking test interchangeability and results have consistently shown that test scores are often highly comparable. This apparent degree of equivalence is a primary contributor to the ongoing development and continued use of CB speaking tests, albeit the issue of absolute versus adjacent currently remains problematic and requires further clarification.

The results from the exemplified case study showed that spoken output throughout test takers' performances was highly comparable between delivery modes. Evidence revealed that differences in the elicited spoken discourse from the test performances is a result of the two different contexts in which the language is elicited. This discourse differentiation manifested itself through the participants' increased output interactiveness via contextualization of the test event in the direct OPI that led to a slight shift in output register akin to the spoken and less literate.

Today, computer-based speaking tests appear to be a feasible alternative to traditional face-to-face oral proficiency interviews. Evidence testifying to this semi-direct delivery mode's reliability and concurrent validity with OPIs, along with the relatively high degree of similarity in test taker elicited discourse provides a stable base from which to encourage the development and use of semi-direct speaking tests. Just as the higher degree of practicality achievable in semi-direct speaking tests cannot be a singular justification for their continued development and implementation, a somewhat higher degree of face validity for direct OPIs should perhaps not be the primary rationale for their ongoing use.

REFERENCES

Abadin, H., Fried, D., Good, G., & Surface, E. (2012). *Reliability study of the ACTFL OPI in Chinese, Portuguese, Russian, Spanish, German, and English for the ACE review*. Raleigh, NC: SWA Consulting.

Alotumi, M. (2018). The effect of CALL-based instruction on students' score attainment on the TOEFL iBT in a Yemeni context. *International Journal of Computer-Assisted Language Learning and Teaching*, *8*(1), 50–64. doi:10.4018/IJCALLT.2018010104

Amengual-Pizarro, M., & Garcia-Laborda, J. (2017). Analysing test-takers' views on a computer-based speaking test. *Profile: Issues in Teachers' Professional Development*, *19*(1), 23–38.

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

Beltran, J. (2016). The effects of visual input on scoring a speaking achievement test. *Working papers in TESOL & Applied Linguistics, 16*(2), 1-23. doi:10.7916/D8543VDT

Berry, V., Nakatsuhara, F., Inoue, C., & Galaczi, E. (2018). *Exploring the use of video-conferencing technology to deliver the IELTS speaking test: Phase 3 technical trial*. Retrieved from https://www.ielts.org/-/media/research-reports/ielts-research-partner-paper-4.ashx

British Council. (2018). *What is Aptis.* Retrieved from https://www.britishcouncil.my/exam/aptis/what

British Council. (2016). *Aptis general speaking test: Practice version*. Retrieved from https://www.britishcouncil.org/aptis-practice-tests/AptisSpeakingPractice/

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT™ and real-life academic speaking activities. *Language Assessment Quarterly*, *11*(4), 353–373. doi:10.1080/15434303.2014.947532

Cambridge English Language Assessment. (2018). *Test your English - adult learners*. Retrieved from http://www.cambridgeenglish.org/in/test-your-english/adult-learners/

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. doi:10371/93290/1/12

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., & Swinton, S. S. (1980). The test of spoken English as a measure of communicative ability in English-medium instructional settings. *TOEFL Research Report*, *7*, 1–68. doi:10.1002/j.2333-8504.1980.tb01230.x

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL*, *20*(1), 21–34. doi:10.1017/S0958344008000311

Dunlea, J., Spiby, R., Nguyen, Q., Nguyen, H., Nguyen, Y., Nguyen, T., . . . Bui, S. (2018). *Aptis-VSTEP comparability study: Investigating the usage of two EFL tests in the context of higher education in Vietnam* (VS/2018/001). Retrieved from https://www.britishcouncil.org/sites/default/files/aptis-vstep_study.pdf

Fairbairn, J., & Dunlea, J. (2017). *Speaking and writing rating scales revision technical report* (TR/2017/001). Retrieved from https://www.britishcouncil.org/exam/aptis/research/publications/scale-revision

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. doi:10.1093/applin/21.3.354

Fulcher, G. (2014). *Testing second language speaking*. Oxon, UK: Routledge. doi:10.4324/9781315837376

Halliday, M. A. K. (1989). *Spoken and written language* (2nd ed.). Oxford, UK: Oxford University Press.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology*, *5*(2), 60.

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. doi: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. doi:10.1080/09588221.2011.649482

Laborda, J. G. (2010). Contextual clues in semi-direct interviews for computer-assisted language testing. *Procedia - Social and Behavioral Sciences 2*. 3591-3595. doi:10.1016/j.sbspro.2010.03.557

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, *13*(2), 151–172. doi:10.1177/026553229601300202

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

Loranc-Paszylk, B. (2015). Videoconferencing as a tool for developing speaking skills. In M. Pawlak, & E. Waniek-Klimczak (Eds.), *Issues in teaching, learning, and testing speaking in a second language* (pp. 189–203). London, UK: Springer. doi:10.1007/978-3-642-38339-7_12

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

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

Mousavi, S. A. (2009). Multimedia as a test method facet in oral proficiency tests. *International Journal of Pedagogies & Learning*, *5*(1), 37–48. doi:10.5172/ijpl.5.1.37

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

Nakatsuhara, F., Inoue, C., & Taylor, L. (2017b). *An investigation into double-marking methods: comparing live, audio and video rating of performance on the IELTS speaking test*. Retrieved from https://www.ielts.org/teaching-and-research/research-reports/ielts_online_rr_2017-1

Ockey, G., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, *32*(1), 39–62. doi:10.1177/0265532214538014

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

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, *28*(3), 373–386. doi:10.1016/S0346-251X(00)00018-X

O'Sullivan, B., & Dunlea, J. (2015). *Aptis general technical manual: Version 1* (TR/2015/005). Retrieved from https://www.britishcouncil.org/aptis-general-technical-manual-version-10

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*(2), 113–125. doi:10.1080/15434300902800059

Quaid, E. D. (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. doi:10.4018/IJCALLT.2018040105

Reemann, E., Alas, E., & Liiv, S. (2013). Interviewer behaviour during oral proficiency interviews: A gender perspective. *Eesti Rakenduslingvistika Uhingu Aastaraamat*, (9), 209-226. doi:10.5128/ERYa9.14

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge, UK: Cambridge University Press.

Rosenfeld, E., Bernstein, J., & Balogh, J. (2005). Validation of an automatic measurement of Spanish speaking proficiency. *The Journal of the Acoustical Society of America*, *117*(4), 2428–2428. doi:10.1121/1.4786630

Schmidgall, J. E. (2017). *The consistency of TOEIC speaking scores across ratings and tasks* (Research Report No. RR-17-46). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12178

Seedhouse, P., & Morales, S. (2017). *Candidates questioning examiners in the IELTS speaking test: An intervention study*. Retrieved from https://www.ielts.org/-/media/research-reports/ielts_online_rr_2017-5.ashx

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

Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral interviews. Anivan, S. (Ed.), Current developments in language testing (pp. 199–209). Singapore: SEAMEO RELC.

Stansfield, C. W., & Kenyon, D. (1992a). The development and validation of a simulated oral proficiency interview. *Modern Language Journal*, *72*(2), 129–141. doi:10.1111/j.1540-4781.1992.tb01093.x

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

Stricker, L. J., & Attali, Y. (2010). *Test takers' attitudes about the TOEFL iBT™*. ETS TOEFL iBT Report No. iBT-13. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2010.tb02209.x

Surface, E. A., Poncheri, R. M., & Bhavsar, K. S. (2008). *Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers*. Retrieved from https://www.languagetesting.com/research

Surface, E. A., Harman, R. P., Watson, A. M., & Thompson, L. F. (2009). *Are human and computer-administered interviews comparable?* Paper presented at the 24th annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals*, *49*(1), 75–92. doi:10.1111/flan.12178

Weir, C. J. (2005). *Language testing and validation: An evidenced-based approach*. London, UK: Palgrave Macmillan. doi:10.1057/9780230514577

Yonezaki, M. (2016). A comparative analysis of semi-direct speaking testing and direct speaking testing for Japanese EFL learners. *International Journal of Curriculum Development and Practice*, *18*(1), 27–38. doi:10.18993/jcrdaen.18.1_27

Lee, Y.-J. (2007). The multimedia assisted test of English speaking: The SOPI Approach. *Language Assessment Quarterly*, *4*(4), 352–366. doi:10.1080/15434300701533661

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

Zheng, Y., & Zhang, Y. (2017). *Aptis in China: Exploring test-taker perceptions of its test validity and practicality* (AR-G/2017/6). Retrieved from https://www.britishcouncil.org/sites/default/files/ying_layout.pdf

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. doi: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. doi:10.118640468-014-0012-y

ADDITIONAL READING

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511733116

Cheung, Y.-R. (2017). Validation of technology-assisted language tests. In C, A. Chapelle, & S. Sauro (Eds.), The handbook of technology and second language teaching and learning (pp. 332-347). Hoboken, NJ: Wiley Blackwell. doi:10.1002/9781118914069.ch22

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. doi: 10371/93290/1/12

Guoxing, Yu., & Zhang, J. (2017). Computer-Based English Language Testing in China: Present and Future. *Language Assessment Quarterly*, *14*(2), 177–188. doi:10.1080/15434303.2017.1303704

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

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*(2), 113–125. doi:10.1080/15434300902800059

Schmidgall, J. E., & Powers, D. E. (2017). Technology and high-stakes language testing. In C, A. Chapelle, & S. Sauro (Eds.), The handbook of technology and second language teaching and learning (pp. 317-331). Hoboken, NJ: Wiley Blackwell. doi:10.1002/9781118914069.ch21

Valencia Robles, J. (2017). Anxiety in language testing: The APTIS case. *Profile: Issues in Teachers'. Professional Development*, *19*(Suppl. 1), 39–50. doi:10.15446/profile.v19n_sup1.68575

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. doi:10.118640468-014-0012-y

KEY TERMS AND DEFINITIONS

Adjacent Agreement: A rating within one sub-level such as advanced-low and advanced- mid.

Concurrent Validity: The degree of test taker score equivalence between two tests, with one test acting as the criterion behavior.

Construct Irrelevant Variance: The introduction of extraneous, uncontrolled variables not reflected in the constructs tested that affect assessment outcomes.

Face Validity: The degree to which a test appears to measure the identified constructs (knowledge and/or abilities) it purports to measure based on the subjective judgment of test stakeholders.

Lexical Density: A quantitative measure of the relationship between lexical and grammatical items in spoken and written discourse.

Output: In the context of this chapter, output refers to test takers' spoken language elicited from a speaking test.

Prosodic Features: Suprasegmental units and forms that occur when sounds are placed together in connected speech. For example, intonation, pitch, stress and rhythm.

Simulated Oral Proficiency Interview (SOPI): First generation semi-direct speaking test that often used tape-mediated delivery in lieu of computers.

Target Language Use (TLU) Domain: The context or situation(s) where the test taker will be using the language on completion of the test.